De l'information à l'influenceFrom Information to Influence

Article · January 2018

DOI: 10.7202/1061789ar

1 author:

Hervé Le Crosnier
C&F éditions
97 PUBLICATIONS **180** CITATIONS

SEE PROFILE

March 2018

Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature

Prepared for:

Authored by:

Joshua A. Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan

Executive Summary

The following report is intended to provide an overview of the current state of the literature on the relationship between social media; political polarization; and political "disinformation," a term used to encompass a wide range of types of information about politics found online, including "fake news," rumors, deliberately factually incorrect information, inadvertently factually incorrect information, politically slanted information, and "hyperpartisan" news. The review of the literature is provided in six separate sections, each of which can be read individually but that cumulatively are intended to provide an overview of what is known—and unknown—about the relationship between social media, political polarization, and disinformation. The report concludes by identifying key gaps in our understanding of these phenomena and the data that are needed to address them.

Outline

Section I: Introduction

Following a relatively brief period of euphoria about the possibility that social media might usher in a golden age of global democratization, there is now widespread concern in many segments of society—including the media, scholars, the philanthropic community, civil society, and even politicians themselves—that social media may instead be undermining democracy (Tucker et al. 2017). This fear extends not just to new or unstable democracies, which are often prone to democratic backsliding, but also to some of the world's most venerable and established democracies, including the United States. Indeed, in little more than half a decade, we have gone from the *Journal of Democracy* featuring a seminal article on social media entitled "Liberation Technology" (Diamond 2010) to the same journal publishing a piece as part of a forum on the 2016 U.S. elections titled "Can Democracy Survive the Internet?" (Persily 2017).

The purpose of this report is to provide a comprehensive overview of the scholarly literature on the relationship between three factors that may be undermining the quality of democracy: social media usage, political polarization, and the prevalence of "disinformation" online.[1] "Disinformation," in the context of this report, is intended to be a broad category describing the types of information that one could encounter online that could possibly lead to misperceptions about the actual state of the world.[2]

Figure 1 on the next page lays out the nature of these concerns. Of perhaps preeminent importance is the question of whether political polarization and/or disinformation decreases the quality of policymaking in democracies, as well as whether it might decrease the overall quality of democracy itself.[3] Further accentuating the problem is the question of whether both these conditions might be fueling each other. That is, does political polarization make people more vulnerable to disinformation, and, in turn, does the increased prevalence of disinformation lead to greater political polarization? Equally important, however, is the third factor: social media usage, which could also possibly be affecting both political polarization and the prevalence of disinformation online. It is this

---

[1] In a prior Hewlett Foundation report (Born and Edgingnton 2017), the authors describe the "information problem" as consisting of three related issues: disinformation, which is *deliberately propagated* false information; misinformation, which is false information that may be *unintentionally propagated*; or online propaganda, which is potentially factually correct information, but packaged in a way so as to *disparage opposing viewpoints* (i.e., the point is not so much to present information as it is to rally public support). While individual literature reviews will report on studies that focus more explicitly on particular subtypes of the information problem, for the purpose of simplicity in this introductory section we use the term "disinformation" to refer to any type of information one could encounter online that could lead to a factually incorrect view of the political world. This could include the now well-known "fake news" (i.e., news emanating from websites that falsely claim to be news organizations while "publishing" deliberately false stories for the purpose of garnering advertising revenue), but also rumors, factually incorrect information, politically slanted information, and "hyperpartisan" news and information.

[2] As is discussed in Section III, actually settling on definitions for these different terms is an important research need moving forward.

[3] On the figure, we include the term "misperception" along with disinformation because the concern is that it is not just disinformation itself, but also the resulting misperception of the political world caused by disinformation, that has the potential to harm democratic quality and policymaking.

triangle—social media driving political polarization and the prevalence of disinformation, both of which are also accentuating each other and simultaneously potentially undermining democratic quality—that has led to so much concern about the potential impact of social media on democracy.

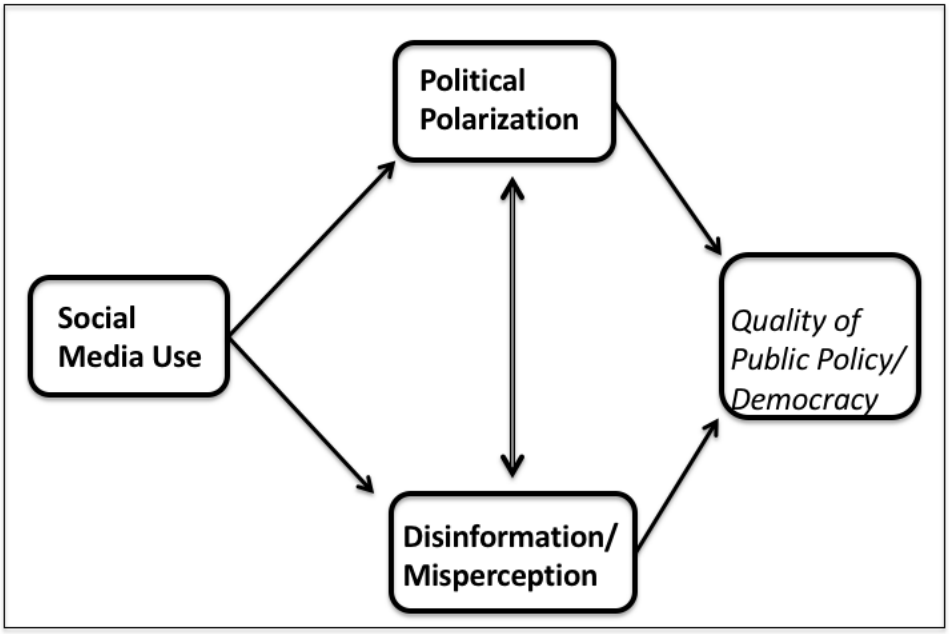

Figure 1. Social Media, Political Polarization, Misperception and Democratic Quality

However, despite our primary interest in these three categories—social media usage, political polarization, and disinformation—there are a number of other related factors of which we need to be aware.

First, there is another path by which we might expect all three of these variables to affect the quality of democracy, which is through political engagement. Social media has been touted as a way of increasing political participation, but it is equally possible that in an era of hyperpartisanship, experiences on social media could also drive people away from politics. Similarly, it might be the case that polarization itself makes politics less attractive for people. Finally, exposure to disinformation may help to mobilize supporters and demobilize opponents (much, we should add, as with many campaign tactics). If we then believe that the quality of democracy is partly a function of the extent to which people are engaged with politics, then all three of these factors could affect democratic quality through impacts on political engagement.

Second, social media, of course, has a complex relationship with traditional media. On the one hand, social media has clearly become a tool for traditional media reporting; one need only think of the number of times a @realDonaldTrump tweet accompanies a news story about the president. At the same time, much of what is shared on social media about politics are stories produced by traditional news media outlets. Further, it seems

increasingly likely that a key goal of online propaganda—often propagated by automated social media accounts, otherwise known as "bots"—is precisely to ensure that some traditional media news stories are viewed more than others (Sanovich et al. 2018).

Finally, politicians themselves have a role to play in this story. They can, of course, create disinformation and/or amplify disinformation from other sources. Elite polarization can increase mass political polarization (Hetherington 2002; Abramowitz & Saunders 2008). Moreover, as recent history has amply illustrated, elites can also play an outsized role in the spread of polarizing content, including through social media. Finally, politicians can intentionally sow distrust in established media orgs to help boost less credible, (possibly social media-based) sources (Ladd 2011). Thus, a more complex model might look like Figure 2:

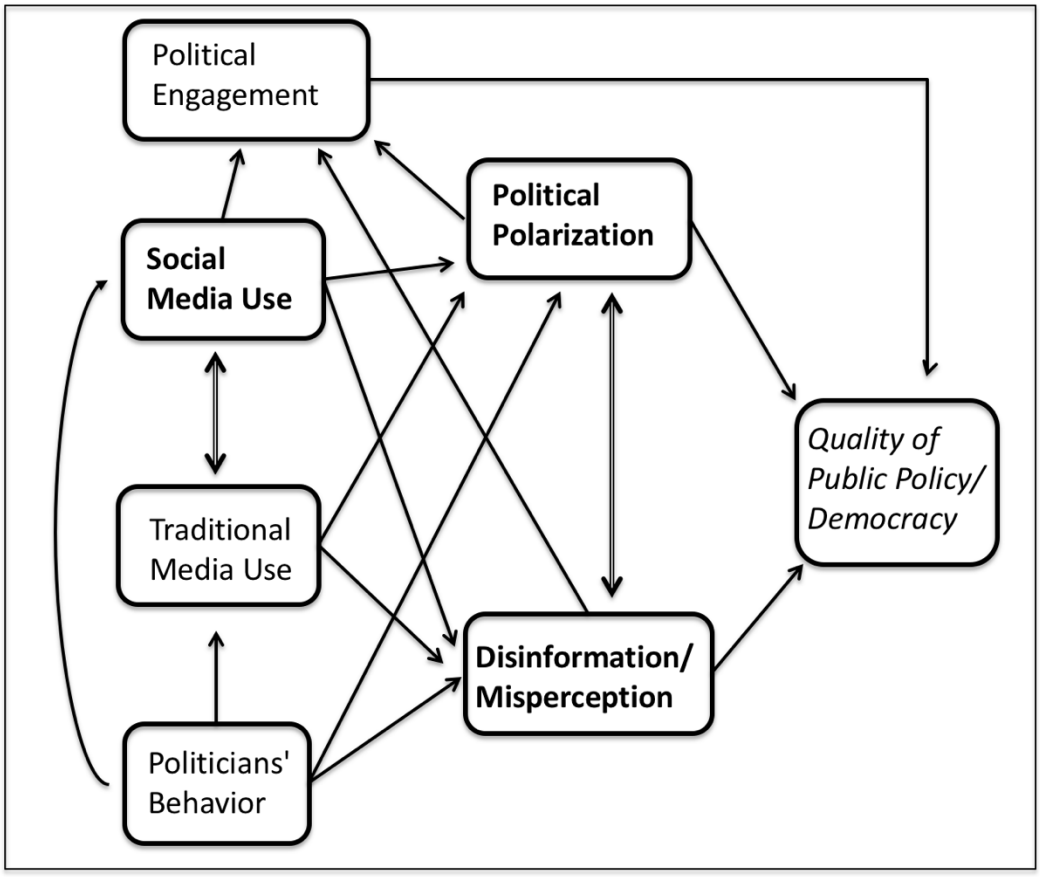

Figure 2. Social Media, Political Polarization, Misperception and Democratic Quality

Two additional points about Figure 2 are worth noting. First, there is no direct arrow linking social media to democratic quality, which is a deliberate choice. While there are many indirect ways in which social media could enhance, or undermine, the quality of democratic governance, we have argued elsewhere (Tucker et al. 2017) that social media itself is neither inherently democratic nor undemocratic, but simply an arena in which

political actors—some which may be democratic and some which may be anti-democratic—can contest for power and influence.

Second, and related, while in the preceding paragraphs we have explained ways in which the various pathways outlined could *undermine* the quality of democratic governance, many of these pathways (with the exception of those flowing through disinformation) could also *enhance* the quality of democratic governance. Indeed, many of the early hopes of the "e-government" movement was that the internet would lead to greater citizen engagement in the monitoring of government actors, as well as greater opportunities for state actors to learn citizen preferences.

Taken together, there are many moving pieces at play in Figure 1, and therefore many questions to untangle as we try to understand whether social media, political polarization, and disinformation are undermining democratic quality, and, if so, how. Fortunately, there is a great deal of scholarly research that has been conducted that can inform how we think about the varied relationships in Figure 1. The purpose of this report, therefore, is to **concisely summarize this research,** in one document, in an effort to allow prospective researchers, philanthropists, civil society organizations, and interested citizens to familiarize themselves with pertinent existing scientific research.

However, we do not currently fully understand all these factors or their relationships to each other. Thus, the second purpose of this report is to **identify key research gaps** in our understanding of the relationships between social media, political polarization, disinformation, and democratic quality. Further complicating matters, even if we can identify the right questions to ask, in many cases we lack the data required for rigorous scientific analyses of these questions. In some cases, the necessary data has simply not yet been collected, but in other cases the necessary data are costly or held by for-profit companies who do not make it available for scholarly research. Thus, the third purpose of this report is to **identify important data needs**.

The rest of the report proceeds as follow. In Section II, literature reviews on six distinct, but interrelated, topics are presented. Each of these reviews was prepared by a separate reviewer (with light editing from the author of the report), and each is preceded by its own executive summary. It is our intention for each of these reviews to function as a stand-alone document that could be read separately by someone interested particularly in that topic, although we want to stress that the topics were chosen because, cumulatively, we hoped they would provide an overview of the current state of the scientific literature on the relationship between our three core variables of social media usage, political polarization, and the spread of disinformation.[4] The six topics are:

A. *Online Political Conversations*
B. *The Consequences of Exposure to Disinformation and Propaganda in Online Settings*
C. *Producers of Disinformation*

[4] Rather than present summaries of each report here, we invite interested readers to see the executive summary at the start of each review.

D. *Strategies and Tactics of Spreading Disinformation through Online Platforms*
E. *Online Content and Political Polarization*
F. *How Misinformation and Polarization Affect American Democracy*

Section III then presents an assessment of the key research gaps in the field cumulatively, across all six topic areas, as well as the data needs for addressing these research gaps in the future.

Research gaps include (1) better estimates of the effects of exposure to information and disinformation online; (2) cross- and multi-platform research; (3) disinformation spread through images and video; (4) the generalizability and comparability of U.S. findings; (5) the role of ideological asymmetries in mediating the effect of exposure to disinformation and polarization; (6) the effects of new laws and regulations intended to limit the spread of disinformation; (7) better understanding of the strengths and weaknesses of different methods of bot detection and analysis; and (8) the role of political elites in spreading disinformation.

Data needs are divided into three categories: data that could be collected in the future by scholars with traditional funding, but that has not yet been collected; data that is prohibitively costly for individual scholars to collect, but that could be provided by a well-funded central research institute/data repository; and data that is not currently available for open scientific analysis due to the fact that it is the property of social media platforms and/or due to privacy concerns.

Finally, Section IV presents a list of all works referenced across all the literature reviews, which we hope will also function as a valuable resource. By definition, the report is intended to concisely summarize broad swaths of academic research; turning to the actual publications that formed the basis of these summaries will in many cases be both recommended and necessary for deeper understanding of the summaries presented here.

Section I: Literature Reviews

Table 1: A Guide to Terms in the Literature Reviews

API	"Application program interface" - means by which platforms allow data to be downloaded.
Bots	Automated accounts that post based on algorithms.
Affective Political Polarization	The extent to which supporters of different political parties dislike the other political party (and possibly its supporters).
Ideological Political Polarization	The extent to which different political parties offer different ideologically distant policy platforms.
Lurkers	People with social media accounts who read posts by others, but do not post themselves.
Social Media Platform	Online architecture for producing content, annotating content produced by others, joining networks to share or view content (e.g., Facebook, Twitter, Instagram).
Social Media Post	Information (text, graphic, video) made available on a social media platform (e.g., a tweet).
Supervised Machine Learning	Machine learning based on training models on labeled outcome data.
Trolls	(1) Human accounts that post politically motivated, generally pro-government content, often for a fee, or (2) human accounts that post provocative (generally "anti-PC") content, often with graphic language and misogynistic content, either out of political conviction or simply for the "thrill" of doing so.
Twitter: mentions	When the name of another Twitter user is contained in a tweet.
Twitter: retweets	When one user shares the tweet of another user.
Twitter: tweets	A "tweet" refers to a post on Twitter; previously limited to 140 characters, recently expanded to 280 characters.
Unsupervised Machine Learning	Machine learning without using a labeled training data set.
VKontakte	Also VK, a Russian social media platform similar to Facebook.

A. Online Political Conversations[5]

Executive Summary

Political conversations, both online and offline, occur most often between people with close personal ties—spouses, close friends, and relatives. The extent to which people are regularly exposed to disagreement, whether via cross-partisan interactions or some other mechanism, remains an open question. This is due to a mix of definitional and methodological issues, combined with a primary focus in the research literature on questions related to the normative ideal of deliberative democracy. This focus has led to studies on the quality of discussion and their effects on outcomes, such as political tolerance and civic engagement. However, more basic questions remain unresolved, such as: How common are informal political discussions on social media? How often do such discussions occur across partisan boundaries? Do these cross-cutting discussions occur primarily via existing relationships or via "weak ties"—for example, friends of friends?

Answering these questions is critical for understanding whether online platforms are contributing to political polarization or serving to dampen its most corrosive effects. As such platforms evolve, researchers should focus on the design features most strongly associated with desirable characteristics of political discussion, such as exposure to cross-cutting perspectives and civility. This section concludes with directions for future research, with suggestions for more use of behavioral data and text analysis.

Studying Political Conversations

There is a rich and varied body of research spanning both political science and communication on the incidence and causes of (mostly face-to-face) political discussion. Following normative concerns about deliberative democracy, much of this research focuses on the quality and follow-on consequences of such discussions: Is political talk civil? Do people engage constructively? Does political discussion lead to greater tolerance? Does it promote civic engagement and political participation (Mutz 2006), or lead to increased levels of knowledge?

In these works, political talk is conceptualized as a central duty of citizenship—as a means of persuading others, resolving conflicts, refining one's own views, and, ultimately, conferring legitimacy upon democratic outcomes. It is therefore not surprising that one of the central preoccupations of scholarship on this topic is the extent to which people encounter disagreement in political conversations. However, this is a difficult question to answer because of three fundamental definitional issues. First, what counts as "political"? Second, what counts as "disagreement"? And third, what counts as a conversation in the

[5] Review prepared by Andrew Guess, Assistant Professor of Politics and Public Affairs, Princeton University.

first place? Scholars' answers to these questions have differed and, lacking consensus, the findings in the literature are somewhat inconsistent (Eveland et al. 2011).

Compounding these difficulties are methodological issues surrounding measurement, sampling, and causation. Measuring the incidence or frequency of political talk typically means asking survey-based questions about respondents' discussion partners and the types of conversations they have had in a given time period. Since such discussions can occur spontaneously, they are subject to biases induced by self-reporting of behaviors—such as voting and media use—which typically result in inflated estimates. How to approach sampling in studies of political discussions is also a difficult question. It depends partially on the unit of analysis: Is it the individual respondent or a discussion itself? And, if one chooses to sample respondents, should it be via traditional random sampling or more complex techniques, such as snowball sampling, that are more specifically tailored to characterizing attitudes and behaviors within a social network? Finally, when studying the effects of political discussion, it is critical for research designs to take into account confounding factors—such as homophily in people's social circles or political interest—that could lead to increased levels of both discussion and broad measures of engagement or participation. Not doing so runs the risk of confusing cause and effect.

Overall, the literature to date is overwhelmingly focused on questions originating from the deliberative tradition in political theory. One consequence is that there is less effort on precisely estimating specific quantities of interest, such as the proportion of political conversations that occur across partisan boundaries, or on making rigorous comparisons across platforms or between online and offline political conversations. Still, there is a rich foundation on which to build a forward-looking research program on cross-cutting exposure to political disagreement in online discussion networks.

Offline Political Conversations

Before turning to research on online conversations, it is useful to summarize the state of knowledge on political discussions in general, primarily from studies that focus on face-to-face interactions. Much of this work is either based on representative surveys, such as the American National Election Studies (ANES), or is qualitative in nature, focusing on smaller subsets of people using an ethnographic approach (e.g., Walsh 2004).

How prevalent is political talk? One of the foundational works in the literature on political discussion networks focused on the context of an election campaign in a single American town (Huckfeldt & Sprague 1995). From their survey data, the authors found that roughly two-thirds of respondents said they talked about politics "only once in a while." Comparing the frequency of discussions about political topics to other subjects, one study in the mid-1990s found talking about "the president, the national government, and the Congress" to be more common than talking about religion or events in other countries, but less common than talking about crime or personal/family matters (Wyatt et al. 2000).[6] Here it may be

[6] The authors only report means from their four-point response scale (from "never" to "often"), so it is difficult to say precisely how prevalent political talk is from their data. Discussions about national political

useful to note the distinction in the literature between informal talk (Walsh 2004) and more formalized forms, such as group forums or organized discussions about specific issues. Regarding the latter type of political discussion, a more recent estimate of participation levels from survey data is 25% (Jacobs et al. 2009).

Who is more likely to talk about politics? There are a number of individual-level correlates of talking about politics with others. These include characteristics associated with having more resources available to devote to informing oneself about politics—income, socioeconomic status, and membership in organizations (Jacobs et al. 2009). Furthermore, indicators of political discussion frequency are often used as part of broader indices of political participation. These suggest a strong relationship to measures of general political interest. As with participation in general, moreover, there is evidence of a gender gap in political discussion: Verba et al. (1997) find that men are more likely to say that they "Discuss national politics nearly every day" than women (31% to 20%) and that they enjoy political discussion (36% to 26%).

How much political talk is cross-cutting? Given the measures used, it is often difficult to back out estimates of the proportion of discussions that are cross-cutting (involving political disagreements or discussions across the partisan divide). One study found that no more than a third of respondents said that *everyone* they discuss politics with supported the same presidential candidate as they did (Huckfeldt et al. 2004), suggesting a relatively high degree of political heterogeneity among discussion partners. The likelihood of exposure to disagreement via conversation appears to be related to strength of partisanship, but only when disagreement is defined in terms of the perceived partisanship of those in one's discussion network (Klofstad et al. 2013). The other important predictor of having a cross-cutting political discussion is the degree of closeness; disagreement evidently occurs more often with casual acquaintances than with close friends or spouses (Mutz & Martin 2001).

Online Political Conversations

Online political discussions occur in environments that differ markedly from a typical face-to-face interaction (Ho & McLeod 2008). For instance, there are fewer contextual cues about discussion partners' reactions (see Walther 2011). Some environments offer anonymity, a feature with significant implications for the quality of discussion (Papacharissi 2004). And discussions are often public or semi-public, visible to many others (Wyatt et al. 2000). Online platforms vary in the extent to which their architectures accentuate these channel characteristics. Anonymity is possible on Twitter and Reddit, for example, while Facebook offers more information about users that could serve as contextual cues. Early research on online political discussions was primarily qualitative in nature (e.g., Kushin & Kitchener 2009), but later researchers have employed traditional survey-based methods, as well as social network analysis.

issues generally averaged just above "sometimes" (3.05), and somewhat below the mean overall for all topics (3.13).

How prevalent is political talk? Using a representative survey of online Americans, Wojcieszak and Mutz (2009) estimated that, at least as of 2006, approximately 11% of internet users reported participating in a message board or chat room of any kind in the past year. Of those, about 17% said they participated in political or civic discussion groups online (as compared to 96% who said they participated in discussion groups related to hobbies or interests). Intriguingly, a substantial proportion of respondents said they discussed politics in the non-political groups—25% of those who participated in leisure groups and nearly half of those who participated in professional groups, for example. These results show the importance of not narrowly conceptualizing political talk as only occurring in designated spaces. They also illustrate a persistent issue with this and related research literatures: Given the pace of change in the online discussion environment, high-quality studies are often obsolete by the time they are published.

Who is more likely to talk about politics? We know somewhat less about this question in the online context due to the constantly evolving nature of both social platforms and online audiences. At a minimum, it appears safe to say that some of the individual-level predictors are similar to those of offline political talk, such as gender, education, socioeconomic status, and political interest (Davis 2005). Moreover, exploratory work on convenience samples has identified traits that could be associated with a *lower* likelihood of talking about politics: conflict avoidance and ambivalence (Jang et al. 2014). These traits may be related to "lurking," or passively following political discussions without necessarily participating (Davis 2005). Evidence suggests that "lurkers" may be more like average Americans than those who actively engage in discussions. (This is an important point to remember when designing and interpreting studies that analyze publicly available social media posts, which select on this trait of active engagement.)

How much political talk is cross-cutting? The answer to this question depends on how one defines disagreement. By simply asking respondents about the level of disagreement (rather than inferring it), Wojcieszak and Mutz (2009) estimate the proportion of discussion groups that expose respondents to cross-cutting arguments or information. The proportion varies based on the type of group, but in general the level of agreement is much higher than the level of disagreement. More than half of political groups primarily exposed respondents to agreement, while about 10% exposed them to disagreement. Other studies have taken different approaches to answering related questions. For example, using an ethnographic approach, one study found a high degree of perceived disagreement in the content of online political discussions (Stromer-Galley 2003), with participants expressing their enjoyment of encountering diverse viewpoints.

While those findings largely represent a mainly web-based discussion environment, later work has focused on political interactions on blogs and Twitter. An influential study of political blogs found a high degree of polarization in the linking patterns of liberal and conservative blogs (Adamic & Glance 2005). That may or may not map onto the concept of political discussion, but studies of Twitter mentions and retweets come closer. One early study of Twitter political interactions has been commonly cited for its finding of strongly polarized retweet patterns within political discussions, shown by a high degree of clustering by the ideological lean of users (Conover et al. 2011). However, the same study

also found much less evidence of such clustering in mention networks. Taken together, these findings suggest that the structure of interaction, imposed by features of the medium itself, can inform the patterns of cross-cutting exposure and polarization that are observed. Even within the same platform, different functions foster vastly different levels of cross-cutting interaction.

Taking this a step further, a recent study of retweet networks across multiple domains found that politically salient topics often resemble "echo chambers" with high polarization (Barberá et al. 2015). However, other topics, such as the Olympics or Super Bowl, more closely resemble "national conversations." It is possible that the best way to achieve cross-cutting exposure in political discussions is via inadvertent exposure within non-political discussion contexts (see also Brundidge 2010). Finally, there are promising innovations in the design of online discussion forums that could encourage greater engagement with cross-cutting comments; in particular, a "respect" (as opposed to "like" option) may have increased interaction with counter-attitudinal comments (Stroud et al. 2017).

Research findings concerning online and offline political conversations exist largely in isolation from each other, although there are exceptions: Stromer-Galley (2002) uses an analysis of data on monthly electronic discussions of political issues to argue that the internet "may provide a new context for political conversation for those who would not normally engage in face-to-face political conversations, thus bringing new voices into the public sphere."

What is the quality of online political talk? An important question related to discussion quality and political polarization is the extent to which online conversations are civil (Papacharissi 2004). While this is a cause for concern, it is unclear how much of online discourse is actually uncivil (even though the most visible interactions may not always be). One recent study of climate change discussions on Twitter found relatively few instances of incivility and sarcasm (Anderson & Huntington 2017). However, a comprehensive analysis of Reddit found a marked increase in incivility there since 2016 (Nithyanand et al. 2017). The authors of that study additionally found greater incivility on Republican subreddits than Democratic ones. They argue that the rise of Donald Trump may have contributed to the increase. Another study focuses on *New York Times* comment threads, finding that incivility can sometimes boost the popularity of comments, despite the preferences of moderators (Muddiman & Stroud 2017). A related strand of recent research has sought to understand the effectiveness of interventions designed to reduce incivility and other normatively undesirable features of online political talk (Munger N.d.). Promising avenues for such interventions focus on the effects of anonymity and social identity.

Directions for Future Research

While the research discussed here has already shed a great deal of light on the nature and prevalence of online political discussions—and how they differ from offline discussions—there is much left to learn. Partially, this is due to the ever-evolving nature of the object of study: Platforms are constantly changing their algorithms and business models, with effects on user behavior that can sometimes be large. Also, while much previous research

has focused on deliberation and the effects of discussion on broader measures of political engagement, there is arguably a need to focus on more grounded questions of the prevalence and types of political discussions that occur online and across different social media platforms. This will foster productive scholarship on the extent of cross-cutting exposure online, the causes and consequences of incivility, and the channel characteristics that encourage or discourage particular forms of political expression.

Methodologically, the field has much to gain from studies that take advantage of large datasets spanning the entire population of potentially relevant discussions rather than relying on inconsistent survey-based reports. This can help to answer questions about overall prevalence. A second area with methodological potential is the use of network approaches (e.g., González-Bailón et al. 2010), which can help clarify the conditions under which strong versus weak ties are important for determining the amount of cross-cutting exposure in online political interactions. Finally, experiments are a promising avenue for testing the *effects* of different types of discussion dynamics on outcomes related to political polarization. Given the growing awareness of affective polarization as a force in American society, it is crucial to identify the mechanisms driving it in as rigorous a way as possible.

B. The Consequences of Exposure to Disinformation and Propaganda in Online Settings[7]

Executive Summary

The spread of political misinformation and propaganda in online settings is generally considered to have negative societal consequences. The conventional wisdom is that "fake news" is amplified in partisan communities of like-minded individuals, where they go unchallenged due to ranking algorithms that filter out any dissenting voice (Pariser 2011). The outcome of this process is a society that is increasingly misinformed and polarized along partisan lines (Sunstein 2017). However, results from empirical studies challenge the different components of this argument: Exposure to political disagreement on social media appears to be high (Bakshy et al. 2015; Duggan & Smith 2016), internet access and social media usage are not correlated with increases in polarization (Boxell et al. 2017), and misinformation appears to have only limited effects on citizens' levels of political knowledge (Allcott & Gentzkow, 2017).

To help address this gap between theory and empirics, we summarize research on three mechanisms by which internet and social media usage may be impacting key societal outcomes of interest. (1) Increased media fragmentation in the online news environment allows citizens to replace political news with entertainment, and lowers the overall quality of the political information being consumed, which limits its potential to increase political knowledge. (2) The consumption of political information through social media increases cross-cutting exposure, which has a range of positive effects on civic engagement, political moderation, and the quality of democratic politics, but also facilitates the spread of misinformation. (3) Political exchanges on social media sites are frequently negative and uncivil, which contributes to the rise in affective polarization.

Introduction

Over the past few years, concerns about the negative societal consequences of the online spread of misinformation and propaganda have become widespread. New technological tools that allow anyone to easily broadcast political information to large numbers of citizens can lead to a more pluralistic public debate, but they can also give a platform to extremist voices and actors seeking to manipulate the political agenda in their own financial or political interest (Tucker et al. 2017). Attention to this problem spiked after the 2016 U.S. presidential election, during which "fake news" was widely shared on social media and reached large numbers of citizens, propagated at least in part by foreign actors (see e.g., Shane 2017). Although there is broad scholarly agreement regarding the high prevalence of misinformation and propaganda in online platforms, whether or not it has

[7] Review prepared by Pablo Barberá, Assistant Professor of Computational Social Science, London School of Economics.

any impact on political outcomes such as levels of political knowledge, trust in democratic institutions, or political polarization remains an open question.

The current conventional wisdom on the impact of misinformation is mostly based on journalistic reports documenting its spread during the 2016 election. Some of the earliest reporting on this topic was produced by Craig Silverman of Buzzfeed News. In a series of articles published around the time of the election, he demonstrated that engagement on Facebook was higher for fake content than for stories from major news outlets. Additional reporting by other outlets corroborated these initial findings (see e.g., Higgins et al. 2016; Rogers & Bromwich 2016; Timberg 2016). Overall, these reports paint a picture of the online news ecosystem in which misinformation and hyperpartisan stories are shared at rates comparable to news stories by mainstream media outlets, reaching millions of people.

This evidence has provided new fuel to the debate on the internet and social media as ideological echo chambers. The prevailing narrative is that online misinformation is amplified in partisan communities of like-minded individuals, where it goes unchallenged due to ranking algorithms that filter out any dissenting voice (see e.g., Pariser 2011; del Vicario et al. 2016). One of the leading proponents of this view is Cass Sunstein, who in his most recent book, *#Republic*, warns that balkanized online speech markets represent new threats to democracy because they are a breeding ground for informational cascades of "fake news" and conspiracy theories (Sunstein 2017). The outcome of this process, he argues, would be a society that is ill-informed and increasingly segregated and polarized along partisan lines, making political compromise increasingly unlikely.

However, the consensus in the scholarly literature is not as clear as these accounts would suggest. Boxell et al. (2017) show that, even if mass political polarization has grown in recent times, this increase has been largest among citizens least likely to use the internet and social media. Their results reveal that "the internet explains a small share of the recent growth in polarization" (p. 10612). Bakshy et al. (2015) and Barberá (N.d.) find that Facebook and Twitter users are exposed to a surprisingly high level of diverse views. Wojcieszak and Mutz (2009) provide similar evidence of frequent cross-cutting political exchanges in online discussion spaces. Survey data collected by the Pew Research Center (Duggan & Smith 2016) show that most users report being exposed to a variety of viewpoints on social media. Forty percent of social media users across different countries report being exposed to a diverse range of sources, according to data from 2017 Reuters Institute Digital News Report (Newman et al. 2017). Finally, regarding the spread of misinformation, Allcott and Gentzkow (2017) find that even if "fake news" stories were widely shared during the 2016 election, the average American saw, at most, several of them on social media.

Put together, this body of work challenges the conventional wisdom, but in many ways raises more questions than it answers. Even if average cross-cutting exposure is relatively high on average, there may be pockets of individuals who are indeed fully embedded in politically homogeneous communities, for whom online consumption of information could lead to increased extremism. Given the nearly universal presence of journalists on social media, messages shared on these platforms could have indirect effects even among the

offline population. We also know little about the long-term consequences of online news consumption on political disaffection, civic knowledge, political participation, and social capital.

There is a clear need for further research addressing the questions above. In trying to structure the discussion of what is known and not yet known within this research agenda, it is useful to consider three potential mechanisms by which online consumption of political information could be impacting political processes: (1) changes in the volume of information being consumed, (2) the (diversity of) sources of such political content, and (3) how it is framed. The following sections discuss the effect of exposure to (mis)information online in key societal outcomes by focusing on how research on these three mechanisms helps resolve the tension between theory and empirics described above, and informs our knowledge of such broader questions.

Volume of Political (Mis)information

In the digital age, anyone can produce and broadcast content that can reach a global audience. There is more political information being shared than ever before, and ordinary citizens now play an active role in the news ecosystem. Bakshy et al. (2015) report that 13% of posts by Facebook users who report their political ideology are "hard news"—national news, politics, or world affairs. Survey data from the Pew Research Center (Shearer & Gottfried 2017) and the Reuters Digital News Report (Newman et al. 2017) shows that two-thirds of Americans, and between 40% and 60% of adults in most developed countries, get news on social media, with Facebook being the leading source. However, in an increasingly fragmented media environment, are citizens still paying attention to politics? Are they better informed?

Several cross-sectional studies report positive correlations between usage of digital media and levels of political knowledge (Baumgartner & Morris 2010; Dalrymple & Scheufele 2007; Groshek & Dimitrova 2011; Kenski & Stroud 2006). However, when interpreting this evidence, we need to be aware that part of these differences could be explained by the online and social media populations being more highly educated and interested in politics.

In an effort to overcome some of the methodological challenges posed by working with cross-sectional data (and self-reported measures of media exposure), Munger et al (N.d.) pair panel survey data with tweets that appeared in respondents' Twitter feeds during the run-up to the 2015 U.K. parliamentary elections. The authors find evidence that tweets from media sources did indeed lead to an increase in knowledge of politically relevant facts, and that exposure to tweets from political parties increased knowledge of the relative placement of parties on different political issues. However, the authors also show that exposure to partisan tweets shifted voters' assessments of the economy and immigration in directions favorable to the parties' platforms—and that much of this movement was in an inaccurate direction—a development more consistent with the expectation of those worrying about pernicious effects from disinformation on social media.

Additional important evidence regarding these questions comes from two field experiments conducted by Theocharis and Lowe (2016) and Foos et al. (N.d.). Both studies randomly assigned access to social media platforms and measured how the use of these tools affected levels of civic engagement. Although political knowledge is only measured here indirectly, the results are similar: The effects of exposure to information are small or even negative. This pattern is consistent with evidence from a panel survey fielded by Dimitrova et al. (2014) and a quasi-experimental survey design in Bode (2016a), which show that digital media use has a limited causal effect on political learning and knowledge.

One potential explanation for this unexpectedly small effect of news consumption is that, even if the volume of political information that is available online is greater, citizens might be tuning out from such content and focusing their attention on entertainment news instead. As Prior (2005) argues, increased media choice could have the unintended consequence of widening gaps in political knowledge: Citizens who are interested in politics increase their news consumption, while those who prefer entertainment become less likely to learn about politics. However, it is still unclear whether this argument applies to social media platforms, where opportunities for chance encounters with political content increase (Fletcher & Nielsen 2017), as discussed in the following section.

Another plausible mechanism is that, even if the overall volume of political information is greater, its average quality is lower. Digital publishing tools have dramatically reduced the costs of producing news, and as a result a large number of new outlets have flourished. The content they produce ranges from high-quality investigative journalism to information that is completely false and misleading, in some cases sponsored by state actors and artificially amplified by bots and other automated accounts (see Reports 3 and 4 below). And, even more complex from a research perspective, there is a wide gray area between these two extremes, which includes clickbait stories, outlets promoting conspiracy theories, hyperpartisan sites, and websites whose business models rely on plagiarizing mainstream media stories (see Review 3). These sites often receive traffic volumes higher than traditional news sites, with social media being an important source of traffic (see e.g., Thompson 2013; Lytvynenko & Silverman 2017). Despite their growing importance in citizens' media diet, we still know little about how consuming this type of (mis)information affects citizens' levels of political knowledge.

Sources of Political (Mis)information

Traditional news consumption is driven in large part by citizens' preference to be selectively exposed to information that aligns with their political views. In contrast, the stories that citizens see on social media are mostly dictated by their social ties. When users navigate these sites, they are exposed to news presented with social endorsements, which affect their probability of reading such content (Bakshy et al. 2012). As Messing and Westwood (2014) show in a series of lab experiments, the presence of social cues reduces partisan selective exposure to levels indistinguishable from chance.

This increasingly *social consumption* of information has a profound impact on societal outcomes, which we are only starting to understand. It likely has a normatively desirable

impact on democratic politics. Studies of the composition of online networks have shown that cross-cutting exposure to information on social media is higher than in offline communication networks or traditional media consumption. Bakshy et al. (2015) show that 20% of the friendships that the average U.S. Facebook users maintains are ideologically dissonant—e.g., 20% of a conservative user's friends are liberal. Barberá et al. (2015) discovered that cross-ideological political interactions on Twitter are more frequent than commonly assumed. Consequently, it is not surprising that Barnidge (2017) finds higher rates of exposure to political disagreement on social media than in face-to-face interactions and more general web browsing.

Political exchanges in such heterogeneous networks have a range of potentially beneficial consequences for democratic citizens. They open up new spaces for civic talk to take place across partisan lines and increase exposure to dissimilar views, which is considered a "central element of the kind of political dialogue that is needed to maintain a democratic citizenry" (Mutz 2006, p.84). And because political elites are also present and active on social media platforms, it could bring politics closer to citizens and make it more transparent, increasing their trust in democratic institutions. Group discussion in diverse online networks may also have positive effects on news seeking and civic engagement (Klofstad 2009; Levendusky et al. 2016; Levitan & Wronski 2014). Cross-cutting exposure could lead to higher levels of political tolerance and awareness of the legitimacy of oppositional viewpoints as well (Mutz 2002). However, not all these effects might be desirable from a normative point of view. As discussed in the following section, cross-cutting exposure may be one explanation behind the recent rise in affective polarization (Suhay et al. 2018).

In contrast with this optimistic view, one could also make a case for a more pernicious impact of the social consumption of news on the health of democratic politics. In a context in which anyone has the potential to make content go viral, journalists' gatekeeping role is diminished, and citizens are likely to be exposed to a larger volume of misinformation and propaganda. Two studies of social fact-checking on Twitter found that citizens' attempt to debunk rumors are generally ineffective (Margolin et al. 2017; Shin et al. 2017). Similarly, Guess et al. (N.d.) and Friggeri et al. (2014) revealed that social fact-checking on Facebook was rare and generally unsuccessful—even if it slowed down the spread of misinformation, it did not stop its propagation, which suggests that ordinary citizens cannot take over journalists' news curation role.

Content and Framing of Political (Mis)information

Empirical studies of exposure to political information on social media reveal an interesting paradox: Most users are embedded in diverse social networks where moderation is the norm, and yet a large share of the content they consume is ideologically extreme and framed in a negative way. This explanation may be behind contradictory findings regarding the effects of the internet on political polarization.

On one hand, Fletcher and Nielsen (2017) find that people who use social networks are exposed to diverse news at a greater rate than people who do not use social networks. This

is not surprising if we consider that a majority of ties in any user's personal network are weak—acquaintances, co-workers, distant relatives, etc. Weak ties play a key role in information diffusion on social media (Bakshy et al. 2012). They are important because of their contribution to the spread of novel information (Granovetter 1973), which is more likely to be ideologically diverse. It is these cross-cutting interactions that have been suggested as a potential mechanism explaining why social media usage does not appear to be correlated with increases in ideological polarization (Boxell et al. 2017).

However, not all users are equally active on social media, and differences in content production across users may help us understand why most political information shared on social media is partisan or extremist. Barberá and Rivero (2015) and Preotiuc-Pietro et al. (2017) show that Twitter users with more extreme ideological positions share disproportionally more content than moderate users. Wojcieszak (2010) finds that extremism increases with frequency of online participation in neo-Nazi online discussion forums. Shore et al. (N.d.) find that even if a majority of Twitter users post links to more moderate news sources than the ones they receive in their own feed, a small core of users does share more extreme content, and they are responsible for the majority of tweets being published. Bakshy et al. (2015) show that the most frequently shared links on Facebook are clearly aligned with largely liberal or conservative populations. Similarly, Flaxman et al. (2016) use web-browsing histories to demonstrate that social media users have higher levels of cross-cutting exposure than those visiting political websites directly, but at the same time they also show higher levels of political segregation in news consumptions.

While this set of studies focuses on the political views being shared on social media, a different dimension may be as important regarding its potential effect on political polarization: the extent to which social interactions through these platforms are uncivil and negative. According to data survey from the Pew Research Center (Duggan & Smith 2016), most social media users in the U.S. find political interactions on social media with people they disagree with to be stressful and frustrating, in large part because they find them less respectful and uniquely angry. Political actors are a frequent target of incivility and harassment: Barberá et al. (N.d.) estimate that 25% of tweets addressed to members of the U.S. Congress contain offensive and incendiary language, and 59% are critical of the politician or their position.

Increased exposure to uncivil disagreement in online contexts has been linked to a range of undesirable effects. Weeks (2015) shows anger increases partisan evaluations of misinformation leading to inaccurate beliefs. Theocharis et al. (2016) find that incivility targeted to politicians makes them less likely to adopt an engaging style, which reduces social media's potential for open, interactive political deliberation. Bode (2016b) presents evidence that political disagreements lead to "unfriending" behavior on social media.

Most importantly, the vitriolic nature of online interactions is likely to be one of the factors explaining the recent rise in affective polarization (Iyengar et al. 2012; Lelkes 2016). As Iyengar et al. (2012) explain, exposure to negative views of members of the opposing party reinforces biased views of out-partisans and increases the perceived social distance between party groups. Recent work by Suhay et al. (2018) provides the best evidence of

how this argument applies to social media: In two experimental studies that randomized exposure to online partisan criticism, the authors found convincing evidence that partisan criticism that derogates political opponents increases affective polarization. This set of results helps us reconcile some of the contradictory findings regarding the connection between social media and political polarization—while it may reduce ideological polarization as a result of leading to higher cross-cutting exposure, it simultaneously may also be increasing affective polarization because of the negative nature of these interactions.

3. Producers of Disinformation[8]

Executive Summary

A diverse combination of actors including **trolls, bots, fake-news websites, conspiracy theorists, politicians, highly partisan media outlets, the mainstream media, and foreign governments** are all playing overlapping—and sometimes competing—roles in producing and amplifying disinformation in the modern media ecosystem. Research spanning across many disparate disciplines has explored the motivations and roles of these actors in the creation and spread of fake news. From detailed ethnographies of troll culture and studies leveraging public opinion data on fake news exposure, to state-of-the-art bot-detection algorithms and sentiment analysis, researchers have taken a wide variety of methodological approaches to identifying and examining the behavior of these actors. This review provides an introduction to each of these sets of actors, and then summarizes the state of the current literature on how each contributes to the production of disinformation.

Independent Trolls

Dating back to the early days of the internet, the term "trolls" has been used to describe people who intentionally bait others in order to elicit an emotional response. Trolls post inflammatory messages to sow discord and cause reactions (Phillips 2015). In the U.S. context, trolls are particularly fond of trolling the mainstream media, tricking outlets into reporting fake stories. Trolls seek to trade up the media food chain, often planting stories with local news outlets, where they are unlikely to be adequately fact-checked. These stories may then gain coverage from mid-sized or national news outlets, as they are either promoted or debunked, amplifying the disinformation far beyond its original scope (Marwick & Lewis 2017). Trolls engage in this behavior both for their own entertainment and to highlight the media's hypocrisy and sensationalism. They frequently claim to be apolitical—arguing that their use of shocking (often racist or sexist) content is simply a convenient tool to offend others (Philips 2015; Higgin 2013). For the purpose of this report we call these people "independent trolls," to distinguish them from the more recent phenomenon of paid political trolls (described in the next sub-section).

However, some scholars have argued that trolls engage in what Coleman (2012, p. 115) calls "the politics of spectacle." Along these lines, the "alt-right" movement and the "manosphere"—blogs and forums devoted to men's rights and misogyny—frequently embrace trolling tactics to draw attention to their causes (Marwick & Lewis 2017).

Trolls are, almost by definition, engaging in polarizing behavior as they seek to foment discord and cause emotional distress. For example, by posting racist or sexist content for the purpose of enraging liberals, trolls may feed into narratives about the rise of online

[8] Review prepared by Alexandra Siegel, Ph.D Candidate, Department of Politics and Graduate Research Associated, NYU Social Media and Political Participation (SMaPP) lab.

hate speech in the Trump era, contributing to a climate of fear and affective polarization, and spreading divisive narratives (Higgin 2013; Herring et al. 2002).

To date, studies of independent trolls and their role in the production of disinformation has been almost entirely qualitative (Phillips 2011, 2015; Marwick & Lewis 2017; Coleman 2012). Tracing the processes by which trolls create content, spread it on mainstream social media platforms, and "trick" the mainstream media into amplifying their mischief clearly requires careful qualitative analysis, but could benefit from a more data-driven approach as well. While some computer scientists have attempted to build "troll detector" algorithms to help social media platforms fight internet trolls and curb cyberbullying (Cambria et al. 2010; Xu & Zhu 2010; Kumar et al. 2014; Xu et al. 2012), these methods have not yet been used to study troll behavior.

Hired Trolls

In contrast to individuals who troll for the satisfaction of eliciting an emotional response and highlighting hypocrisy, hired trolls are people who are paid by companies, politicians, political parties, and other actors to write fake posts and comments in public forums (Mihaylov et al. 2015).

For example, media reports and Western intelligence reports suggest the presence of Russian "troll farms," where employees are given quotas and instructed to influence conversations about regional, national, and international issues. These reports suggest that the Russian government is employing trolls as part of conscious strategy to sway public opinion in its favor and against the United States and its NATO allies, both domestically and abroad (Gerber and Zavisca 2016).

To date, almost no published academic research has been devoted to the study of these trolls. While a variety of scholars have written about hired trolls as a component of Russia's broader disinformation strategy,[9] it is difficult to study their behavior systematically given their motivation to go undetected and their employers' motivations to keep their existence secret. Computer scientists are working on developing hired troll detection algorithms based on the frequency with which users on diverse platforms accuse other users of being trolls, although these methods require more human validation (Mihaylov et al. 2015). Recent work has begun to empirically examine the role of hired trolls in spreading disinformation, for example the role of Russian trolls in the #BlackLivesMatter movement (Stewart et al. 2018), as well as the ability of trolls in Russia to change the direction of conversations on blogging platforms (Ananyev & Sobolev 2017).

Bots[10]

Bots are pieces of software that create content on social media (Forelle et al. 2015). A growing body of research is devoted to the study of computational propaganda—the use of

[9] For more details on this strategy, see the "Foreign Governments" section below.

[10] For much more on bots, see Review E (below) on "Strategies and Tactics of Spreading Disinformation."

algorithms, automation, and human curation to purposefully distribute false or misleading information over social media networks.[11] Scholars have uncovered evidence of bots and other forms of computational propaganda in the American social media sphere and in diverse international contexts, including Azerbaijan, Brazil, Canada, China, France, Germany, Italy, Mexico Poland, Russia, Ukraine, and Venezuela (Forelle et al. 2015; Ferrara 2017; Woolley 2016; Bessi & Ferrara 2016; Treré 2016; Ferrara et al. 2016; Shorey & Howard 2016; Kollanyi et al. 2016; Marwick & Lewis 2017; Stukal et al. 2017). Evidence from these studies suggests that bots are used for a variety of deceptive political purposes. These include inflating politicians' follower and "like" counts (Woolley 2017); influencing political discourse (Forelle et al. 2015); attacking dissidents (Treré 2016); manipulating public opinion (Woolley 2016; Kollanyi et al. 2016); and possibly for manipulating news search rankings (Sanovich et al. 2018).

A growing body of research has been devoted to the use of bots in the 2016 U.S. election.[12] For example, Bessi and Ferrara (2016) use bot detection algorithms and data collected with Twitter's streaming API to show that bots account for about one-fifth of tweets about the U.S. 2016 election during the final month of the campaign. Kollanyi et al. (2016) report that bots were quite active in producing pro-Trump, and to a lesser extent pro-Clinton, content during the presidential debates. Bots have additionally been known to engage in harassment and hate speech in political conversations, generally contributing to a climate of polarization and enmity in online political discussions (Kollanyi et al. 2016).

Other research has examined the role of bots in electoral campaigns in Latin America (Ferrara 2017; Suárez-Serrato et al. 2016), the U.K. (Howard & Kollanyi 2016), France (Ferrara 2017), and Italy (Cresci et al. 2017). These studies all point to the role of bots generating a large volume of social media posts to support, or attack, candidates or positions. During the 2016 U.K. Brexit referendum researchers found that political bots played a small but strategic role in disseminating hashtags associated with the "leave" campaign. These bots were prolific, with less than one percent of sampled accounts generating almost a third of all the messages containing "leave"-related hashtags (Howard & Kollanyi 2016). Some studies even suggest that bots are recycled—namely the same bots are used in different electoral campaigns (Ferrara 2017; Starbird et al. 2014; Nied et al. 2017). For example, Ferrara (2017) demonstrates that a series of bots that were producing alt-right narratives during the 2016 election disappeared after November 8, 2016, and then reappeared in the run-up to the 2017 French election, tweeting anti-Macron content.

Fake News Websites

Some actors producing fake news are apparently in it just for the money. Because social media is a largely unregulated medium, supported and driven by advertising, some purveyors of disinformation may be purely profit-maximizing (Burkhardt 2017; Bakir & McStay 2017; Allcott & Gentzkow 2017; Marwick & Lewis 2017). When news articles go

[11] See Woolley and Howard (2017) for an overview.

[12] Although note that the first reports of coordinated attacks against political candidates on social media date back to 2010 (Metaxas & Mustafaraj 2012; Ratkiewicz et al. 2011a; Ratkiewicz et al. 2011b).

viral, they generate advertising revenue each time a user visits the original site where the content appeared, incentivizing diverse entrepreneurial individuals to get involved in the business of fake news (Allcott & Gentzkow 2017).

Along these lines, journalistic investigations suggest that more than 100 sites producing fake news articles during the 2016 U.S. election campaign were run by teenagers in a small town in Macedonia. They created a range of websites with names like USConservativeToday.com and posted stories claiming—among other things—that Hillary Clinton would be indicted for crimes related to her emails (Marwick & Lewis 2017). Stories they produced—favoring both Trump and Clinton—earned them tens of thousands of dollars (Subramanian 2017). Similarly, a U.S.-based fake news producer, Paul Horner, ran a successful fake news site called National Report for years prior to the 2016 election (Dewey 2016).

In general, these actors claim to be apolitical. During the 2016 election, they claimed to be motivated by profit, and publishing pro-Trump content generated more advertising revenue than pro-Clinton content (Marwick & Lewis 2017). However, profit maximization may not be the only motivation behind fake news websites. For example, the Romanian man who ran endingthefed.com, asserts that he started the site mainly to help Donald Trump's campaign (Townsend 2016). By contrast, other producers of right-wing fake news identify as liberal and sought to embarrass those on the right by demonstrating that they would gullibly disseminate false stories (Dewey 2016; Sydell 2016). Regardless of whether or not ideology plays a role, the costs of entering the market and producing fake news content on social media are extremely minimal. As a result, small-scale, short-term strategies adopted by fake news producers can be extremely profitable and these actors have little incentive to develop trusted reputations (Marwick & Lewis 2017).

In one of the only empirical studies of the influence of these fake news sites during elections, Allcot and Gentzkow (2017) demonstrate that fake news was both widely shared in the 2016 campaign period and heavily tilted in favor of Donald Trump. They show that a list of fake news websites, on which just over half of articles appear to be false, received 159 million visits during the month of the election. Using web browsing data and an online survey, they estimate that the average American adult saw and remembered 1.14 fake stories. Regarding the role of these sites in influencing polarization, the authors show that Democrats and Republicans are both about 15% more likely to believe ideologically aligned headlines, and this effect is substantially stronger for users in homogenous social media networks.

Conspiracy Theorists

From amateur filmmakers who post conspiracy "documentaries" on YouTube to 4chan and Reddit users propagating dubious claims, the internet is a fertile breeding ground for conspiracy theories (Clarke 2007; Marwick & Lewis 2017). It has been argued that these corners of the web are particularly likely to become echo chambers, as skeptical users often opt out of these communities (Wood et al. 2012).

Conspiracy theorists often express anxieties about losing control or status. They are driven by a belief that a powerful group of people is manipulating the public, while concealing their activities (Sunstein & Vermeule 2009). These claims range from anti-Semitic conspiracies about Jews plotting to take over the world to alternative accounts of specific events such as the 9/11 attacks or the Sandy Hook school shootings (Marwick & Lewis 2017).

Mass media often amplifies conspiracy theories, profiting off of their appeal. For example, news channels feature "documentaries" investigating such theories without debunking them (Byford 2011). In 2011, when Trump began propagating the "birther" conspiracy theory, asserting that President Obama was born outside the United States, mainstream news outlets covered these claims extensively (Wiggins 2017). A new industry of conspiracy theories, typified by Alex Jones' multimedia franchise that works to spread conspiratorial content, has also emerged.

As Marwick and Lewis (2017) argue, the modern media ecosystem perpetuates a cycle in which online communities rely on conspiracy-driven news sources, whose claims are covered by mainstream news media, thereby exposing the public to these ideas. This phenomenon is strikingly widespread. In fact, survey data suggest that half the American public consistently endorses at least one conspiracy theory, and belief in particular conspiracy theories is often divided along ideological lines (Oliver and Wood 2014).

Conspiracy theories also played a role in the 2016 election campaign. In particular, Donald Trump frequently amplified conspiracy theories, many of which can be directly traced to Alex Jones and his website Infowars (Finnegan 2016, Marwick & Lewis 2017). Nevertheless, the extent to which this type of amplification actually plays a role in people's belief in rumors and, ultimately, electoral choices remains unclear.[13]

Hyperpartisan Media

Faris et al. (2017) argue that highly partisan media is the primary incubator and disseminator of disinformation today. Over the last decade, an extensive network of hyperpartisan right-wing news sites and blogs has emerged (Faris et al. 2017; Eldridge 2017). The American far right has a history of exploiting new media to advance their ideological agenda—from their use of anti-communist radio in the 1950s to the rise of right-wing talk radio in the 1990s (Faris et al. 2107; Marwick & Lewis 2017). This new landscape of hyperpartisan media is dominated by sites such as *Breitbart*, *the Daily Caller, The Gateway Pundit, the Washington Examiner, Infowars, Conservative Treehouse,* and *Truthfeed*.

Faris et al. (2017) characterize these actors as "combining decontextualized truths, repeated falsehoods, and leaps of logic to create a fundamentally misleading view of the world." Such sites frequently spread misinformation, rumors, conspiracy theories, and

[13] For much more on the topic of correction of misperceptions, see Review F on how misinformation and polarization affect American democracy.

attacks on the mainstream media (Marwick & Lewis 2017). Hyperpartisan sites also exist on the left. These include: *Occupy Democrats, Addicting Info, Daily Newsbin*, and *Bipartisan Report*. However, on the left, such sites are far less influential than center-left or mainstream news sites (Faris et al. 2017).

In the most comprehensive empirical study of hyperpartisan media both on and offline, to date, Faris et al. (2017) demonstrate that hyperpartisan news is much more widely shared on Facebook than the types of explicitly fake news stories described in the previous section. Examining linking patterns on news websites, they also find that the U.S. media environment is asymmetrically polarized. The far right is a dense, tightly linked network that is largely isolated from other media sources, whereas the far left is largely integrated into the mainstream media discourse.

Politicians

Politicians themselves are also responsible for producing and amplifying disinformation in a variety of contexts. By sharing information in support of their positions, politicians can gain attention, popularity, and support (Marwick & Lewis 2017). Through attracting large numbers of followers on social media (Vaccari & Valeriani 2015), they become central nodes in online networks. As politicians seek visibility and support, they can (inadvertently or intentionally) produce or amplify the spread of disinformation. A similar process can also occur when politicians disseminate false information through hyperpartisan media channels (Berinsky 2017).

Lying in politics is hardly new (Jay 2010), but qualitative evidence from interviews with journalists and media companies indicates that politicians today make more false claims than ever before—both in the U.S. and in other contexts (Skjeseth 2017). In the US, recent research has focused on the unprecedented frequency with which Donald Trump makes false statements—far outpacing other political candidates during the 2016 election (McGranahan 2017; Skjeseth 2017).

Regarding the role of politicians in the amplification of disinformation, the vast majority of shared content online does not "go viral" or spread in cascades among average people. Accounts from politicians, with high numbers of followers, however, can dramatically increase the reach of information on social media, and their posts are often reported on in the mainstream media. As a result, politicians may not be the largest sharers of disinformation, but they might be some of the most influential (Mele et al. 2017).

Foreign Governments

A great deal of journalistic and scholarly attention has recently been devoted to Russian attempts to spread disinformation and sow discord in the 2016 election. (See Maréchal 2017 for an overview.) But qualitative historical research suggests that foreign

governments have long spread disinformation and propaganda in order to advance their agendas abroad (Mele et al. 2017; Schudson 1997).[14]

Studies of the Russian government's current disinformation strategy are largely qualitative. They often situate Russia's modern strategy within the Soviet-era practice of dezinformatsiya (disinformation), or planting false or distorted stories to influence Western public opinion (Ziegler 2017; Maréchal 2017). In recent years, Russia, China, Iran, and Venezuela have all used various disinformation strategies—successfully and unsuccessfully—to counter Western democracy promotion and to promote authoritarian interests abroad (Vanderhill 2013; Way 2015; Nocetti 2015; Lankina & Watanabe 2018). While Russian efforts were initially focused on the "near abroad" including the Ukraine, Belarus, and Georgia, the strategy now reaches far beyond the former Soviet republics.[15]

Social media and the internet have magnified the impact of this "information warfare" by authoritarian governments (Diamond et al. 2016; Tucker et al. 2017; Roberts 2018, Sanovich et al. 2018). The fragmentation of the current media landscape makes Western channels vulnerable to unwittingly amplifying narratives pushed by state-run media outlets like *Russia Today*. This facilitates authoritarian regimes' manipulation of the perception of key issues by making it more difficult to distinguish between authentic and false information (Diamond et al. 2016; Maréchal 2017; Richey 2017).

The Western intelligence community has also devoted a great deal of resources to understanding Russia's disinformation strategy, though their exact methodological approach is, of course, a black box. Using intelligence information collected by the FBI, NSA, and CIA, a recent report from the Office of the Director of National Intelligence (2017) concludes that Russian President Vladimir Putin ordered an influence campaign in 2016, aimed at the U.S. presidential election, designed to undermine public faith in the U.S. democratic process, denigrate Secretary Clinton, and harm her chances of election. They find that Russia's strategy evolved over the course of the campaign based on Russia's assessment of Clinton and Trump's electoral prospects. When Moscow became convinced that Secretary Clinton was likely to win the election, the Russian influence started to focus more on undermining her future presidency. They find that this strategy combined covert intelligence operations—such as cyber activity—with overt efforts by Russian Government agencies, state-funded media outlets like RT, third-party intermediaries, and internet trolls.

Conclusions

Across these literatures, both the depth of qualitative understanding of these actors and the rigor with which their behavior has been empirically examined are quite uneven. For example, while a wealth of ethnographic work paints a very clear picture of the origins,

[14] For example, the British waged an effective fake news campaign around alleged German atrocities during World War I in order to mobilize domestic and global public opinion against Germany. These efforts, however, had unintended consequences, because memories of that disinformation led to public skepticism during World War II when reports first emerged about Holocaust atrocities (Schudson 1997).

[15] See Ziegler (2017) and Gerber and Zavisca (2016)for an overview of this literature.

motivations, and behaviors of independent trolls, algorithmic approaches to troll-detection are largely lacking, and almost no empirical work has investigated troll behavior systematically on or across internet platforms. By contrast, while researchers have done impressive empirical research on the prevalence and behavior of bots in a variety of electoral contexts, more qualitative, experimental, and mixed-method work is needed to understand how people interact with and perceive bot behavior.

The literature on the producers of disinformation is atomized, and it suffers as a result. Because each of these actors is interacting in a complex media ecosystem, studies that examine the behavior of multiple actors, conduct research across platforms, and integrate online and offline media data sources—a la Faris et al. (2017)—provide us with the most nuanced and policy-relevant understandings of the producers of fake news.[16] This is especially crucial if we wish to understand the true scope of the role these actors are playing in impacting electoral outcomes or driving political polarization.

[16] Marwick and Lewis (2017) provide an impressive account of the complex interplay of many of these actors, and future research on this topic will benefit from their groundwork.

D. Strategies and Tactics of Spreading Disinformation through Online Platforms[17]

Executive summary

This review is divided into three parts. We begin with addressing the primary *tactics* for spreading disinformation online: censorship; hacking and sharing; the manipulation of search rankings; and the use of bots and trolls to directly share information. In the second section, we summarize the current state of the ever growing literature on what we know about how bots and trolls have been employed in the disinformation sphere, as well as providing a short technical discussion of the current state of bot and troll detection techniques. In the final section, we look at some of the underlying characteristics that make social media platforms inherently susceptible to disinformation campaigns, namely the dependence on ad revenue and the use of optimization algorithms.

Tactics for Spreading Disinformation

There are numerous ways in which disinformation can be spread online. In this section, we consider four tactics: selective censorship; the manipulation of search rankings; hacking and releasing; and directly sharing disinformation on social media platforms.

We begin with *selective censorship*, which involves removing some content from online platforms, while leaving other forms content alone; King et al. (2013) document this type of activity using Chinese data. To the extent that this curated approach to removing some content serves to privilege disinformation that is not censored, it is a clear tactic for spreading disinformation.

A second tactic is to try to *manipulate search algorithms* to make certain news stories (or sources of disinformation) more likely to appear, for example, in a Google search. This is not completely dissimilar from normal advertising, as well as spam campaigns (Metaxas 2010). Traditional spamming tactics include *keyword stuffing* (adding popular keywords to promote websites in search engine rankings); *link bombs* (using anchor text in links to relate specific search queries with required websites);[18] and creating *mutual admiration societies* (groups of websites with links pointing to each other).

These techniques have been adopted for the purposes of computational propaganda on social media. In particular, *keyword stuffing* has been used to make posts with predefined keywords/hashtags to promote specific messages; *link bombs* have taken the form of

[17] Review prepared by Sergey Sanovich and Denis Stukal, Ph.D Candidates, Department of Politics and Graduate Research Associates, NYU Social Media and Political Participation (SMaPP) lab.

[18] One example here would the 2006 Google-bomb that would make Google show results related to George W. Bush for the query "miserable failure." The mechanics are based on the anchor text in links: One needs to make a link to a webpage related to G.W. Bush and add the anchor text "miserable failure". Since anchor texts are assumed to be a good description of a webpage, Google would relate that anchor text with the webpage to which the link points.

similar or identical posts pointing to specific websites; *mutual admiration societies* are groups of accounts that follow and repost/retweet each other. Sanovich et al. (2018), in their analysis of a sample of verified bots in Russian Twitter in 2014–2015, find that about 40% of the accounts tweet news headlines *without* links to the news story, 40% tweet headlines *with* a link to the story, and another 10% consist entirely of retweets from other accounts. According to Metaxa-Kakavouli and Torres-Echeverry (2017), during 2016 U.S. presidential and senatorial elections "up to 30% of […] national candidates had their search results affected by potentially fake or biased content."

This type of algorithmic manipulation can also take place within social media platforms, for example with trending topics and hashtags on Twitter and Facebook. The degree of manipulation could serve as an independent tool for measuring disinformation campaign success, particularly valuable for people paying for the campaign, but also for its managers (see Sanovich 2017, and a detailed case study by Fedor & Fredheim 2017).

Within social media platforms, related tools include hijacking hashtags, popular users' mentions, and other venues that serve as focal points for information exchange and action coordination. Such strategies could include cluttering conversations with either counter-messaging delivered in bulk, or even simpler, with some distracting or meaningless content, and could destroy the organizing power of social networks.

A third strategy for disseminating disinformation involves *hacking sensitive and/or damaging information* (primarily from email accounts) and subsequently selectively leaking the information—in either its real form or following manipulation of the hacked materials—so as to damage the targets of disinformation campaigns. By far the most famous example of this type of operation is the alleged hacking of the email accounts of the Democratic National Committee and Hillary Clinton's campaign chairman John Podesta, and the subsequent use of an army of trolls and bots to spread damaging information related (or supposedly related) to these emails, in an effort to impact the 2016 U.S. presidential election. It is not as widely known that the Russian government was following the same playbook in regard to its domestic opposition for at least a decade (Sanovich 2017).

Finally, and perhaps most importantly, disinformation campaigns can be conducted by *directly introducing disinformation* onto social media platforms and then helping to spread that disinformation.

In the previous chapter of this report (Review C), we introduced readers to bots and paid trolls, both of which are categories of actors that can play important roles in manipulating search rankings, sharing hacked information, and directly introducing and sharing disinformation on social media platforms. We therefore turn in the next section specifically to summarizing the existing literature on the role played by bots and trolls in disinformation campaigns. Before doing so, however, it is important to note that there are other potentially important actors in this regard that have not received as much attention in the academic literature. These include *influential bloggers,* recruited for the campaign in exchange for substantive monetary compensation (which is often set above market rate for

advertisement and product placement); and *activists and/or government officials* who don't get paid directly for participation in the campaign, but belong to the political machine behind it. Finally, a campaign could also benefit from the involvement of the *ordinary users*, who are mobilized by the people in the above categories to support the same cause on a volunteer basis.

Bots and Trolls: Scale of Activity and Impact

The use of bots and trolls in disinformation campaigns around the globe is by now well-documented.[19] Bessi and Ferrara (2016) identified 400,000 bots responsible for posting about 3.8 million tweets during the last month of 2016 U.S. presidential elections. This constituted about one-fifth of the total volume of online conversations they had collected. Shao et al. (2017) expose the role of bots in spreading fake news, especially at the early stages of the dissemination. They also note that "humans are vulnerable to this manipulation, retweeting bots who post false news bots." Reports from, among others, NBC News, *WIRED*, *Wall Street Journal*, and CNN, give ample qualitative evidence of bot and troll activity, including setting up Facebook groups; attempting to organize offline events; and spreading highly explosive and divisive messages on racial relations, gun and abortion rights, etc.[20]

Beyond the 2016 U.S. presidential election, Russian online disinformation operations were, at least allegedly, uncovered during other electoral campaigns in the United States as well as in the post-election period.[21] This includes the German (Applebaum et al 2017); and

---

[19] Estimates of the total number of bots on Twitter vary a lot. The official Twitter estimates released in 2016 claim that around 8.5% of Twitter users are bots. Varol et al. (2017) estimate that between 9% and 15% of active Twitter accounts are bots. Wei et al. (2015), on the other hand, estimated that 50% of the Twitter accounts created in 2014 were bots. But for the purposes of this review, only political bots are of relevance and they, ideally, need to be measured against the total volume of *political* discussion. As we edit the final version of this report, yet another attempt has been made by Twitter to remove bots following the tragic Parkland, Florida, school shooting and the subsequent online #crisisactors disinformation campaign being waged against students from Stoneham Douglas High School. See Sacks, Brianna. February 21, 2018. "Nope, The Florida School Shooting Survivors Demanding Gun Control Are Not Crisis Actors." *BuzzFeed*. https://www.buzzfeed.com/briannasacks/nope-the-florida-school-shooting-survivors-demanding-gun; Ashley O'Brien, Sara. February 21, 2018. "Twitter is trying to crack down on spam bots." *CNN Money.* http://money.cnn.com/2018/02/21/technology/twitter-lockout/index.html.

[20] O'Sullivan, Donie, and Dylan Byers. September 28, 2017. "Fake Black Activist Social Media Accounts Linked to Russian Government." *CNN Money*. http://money.cnn.com/2017/09/28/media/blacktivist-russia-facebook-twitter/index.html; Parham, Jason. October 18, 2017. "Russians Posing as Black Activists on Facebook Is More Than Fake News." *WIRED*. https://www.wired.com/story/russian-black-activist-facebook-accounts/; Popken, Ben. November 30, 2017. "Russian Trolls' Graphic Tweets on Racism, Rape, and Satanism Revealed." *NBC News.* https://www.nbcnews.com/tech/social-media/russian-trolls-pushed-graphic-racist-tweets-american-voters-n823001; Wells, Georgia, and Deepa Seetharaman. October 13, 2017. "Facebook Users Were Unwitting Targets of Russia-Backed Scheme." *Wall Street Journal*. https://www.wsj.com/articles/facebook-users-were-unwitting-targets-of-russia-backed-scheme-1507918659.

[21] Clifton, Denise. December 11, 2017. "Russian Propagandists Are Pushing for Roy Moore to Win." *Mother Jones*. http://www.motherjones.com/politics/2017/12/russian-propagandists-are-pushing-for-roy-moore-to-win/.

British[22] (Gorodnichenko et al. 2017) elections, and the Catalonian[23] referenda. Bots detected during the so-called *MacronLeaks* hoax immediately before the French presidential elections in 2017 (Ferrara 2017) were variously attributed to both Russian and American supporters of Macron's main opponent Marine Le Pen.[24] In addition, NATO conducted a study of Twitter discussions regarding its presence in Baltic countries and Poland (Fredheim 2017). They claim that 70% of accounts tweeting in Russian during five months in 2017 appeared to be automated (as opposed to 28% for accounts tweeting in English, see Fredheim 2017).

Studies of Russian domestic politics, as well as its conflicts with neighbors, reveal no less evidence of active deployment of bots and trolls for propaganda purposes. Stukal et al. (2017) demonstrate that between February 2014 and December 2015 (an especially consequential period in Russian politics that included the annexation of Crimea and Russian involvement in the conflict in Eastern Ukraine), on the majority of days, the proportion of tweets in their collection (tweets from primarily Russian language accounts, selected based on keywords related to Russian politics, meeting a minimum threshold of activity) produced by bots exceeded 50% of the total volume of tweets in the collection. Ananyev and Sobolev (2017) provide causal evidence of confirmed Russian government trolls being able to change the direction of conversations on the LiveJournal blogging platform that was popular in Russia in the 2000s and early 2010s. Labzina (2017) shows that Russian astroturfing extended even to Wikipedia, where trolls from the infamous Russian "troll factory" *Internet Research Agency* (identified by IP geographic location) [25] made contributions to Wikipedia articles in support of the Russian government's political positions and historical narratives.

China is another major actor whose online activity, including bots and trolls, is actively investigated by researchers. King et al. (2017) show how the Chinese "50-centers" are used to provide a positive distraction from discussions of any controversial issues. Miller (2017) shows that as much as 15% of all comments made on 19 popular news websites in China are made by government astroturfers. Tsay (2017) shows how Chinese government astroturfing works in practice, using data about official police accounts on Sina Weibo.

While China and Russia are the focus of the bulk of research so far, there is also research on

---

[22] Booth, Robert, Matthew Weaver, Alex Hern, and Shaun Walker. November 14, 2017. "Russia Used Hundreds of Fake Accounts to Tweet about Brexit, Data Shows." *The Guardian*. http://www.theguardian.com/world/2017/nov/14/how-400-russia-run-fake-accounts-posted-bogus-brexit-tweets.

[23] Alandete, David. November 11, 2017. "Russian Network Used Venezuelan Accounts to Deepen Catalan Crisis." *EL PAÍS*. https://elpais.com/elpais/2017/11/11/inenglish/1510395422_468026.html.

[24] Hern, Alex. May 8, 2017. "Macron Hackers Linked to Russian-Affiliated Group behind US Attack." *The Guardian*. http://www.theguardian.com/world/2017/may/08/macron-hackers-linked-to-russian-affiliated-group-behind-us-attack; Politi, Daniel. May 6, 2017. "American Alt-Right and Twitter Bots Are Key to Spreading French Election Hack." *Slate*. http://www.slate.com/blogs/the_slatest/2017/05/06/american_alt_right_and_twitter_bots_are_key_to_spreading_french_election.html.

[25] Chen, Adrian. June 2, 2015. "The Agency." The New York Times. https://www.nytimes.com/2015/06/07/magazine/the-agency.html.

bot activity in Japan and South Korea (Schäfer et al. 2017; Keller et al. 2017) and the impact of suspended Twitter accounts (at least part of which, presumably, were bots) on Arabic Twitter during the Arab Spring (Wei et al. 2015). In the Japanese and Korean cases, bots were operating on behalf of the incumbents (the same leader in Japan, and the same party in Korea where the previous leader was term-limited) at a time when they were facing competitive reelection campaigns. In the Japanese case bots were likely operated privately by political allies of the incumbent; in Korea they were operated as part of a secret intelligence operation intended to ensure the incumbent party's candidate reelection. Keller et al. (2017) find a very limited impact for bots on human accounts' behavior, while Wei et al. (2015) show that suspended Twitter accounts had a substantial effect on hashtag rankings during the Arab Spring, although this had little effect on the distribution of topics.

Micro-level evidence so far is more consistent in finding a significant impact of bots and trolls on individual human users' perception and behavior. One important finding in that direction is that human users do not necessarily have less trust in bot accounts than in human ones. Edwards et al. (2014) conducted an experiment that involved two groups of students inspecting identical Centers for Disease Control Twitter pages that differed only in that one stated explicitly that it was a CDC Twitter bot, whereas the other said that it was a CDC researcher. The participants then responded to a number of questions measuring their perception of the credibility, interpersonal attraction, and communication competence of the inspected accounts. The study found statistically significant differences in social and task attraction, but not in credibility and competence. This, however, was arguably not an instance when respondents would have expected the bot to be engaged in providing disinformation.

Further research is required to find the extent to which these results generalize to other types of Twitter bots (in particular, those that try to hide their non-human nature) and across substantive domains (including politics).

The need for further research on the human perception of bots in different substantive domains is partly justified by scholarly findings that show differing human abilities to identify bots in different spheres. Everett et al. (2016) mixed a large number of Reddit comments written by humans with automatically generated texts from a second-order Hidden Markov Model on five different topics (including science and adult topics) and different crowd sentiment.[26] They then recruited two panels of coders (three cybersecurity researchers and three typical internet users who browse social media on a daily basis) and asked them to label comments as written by humans or bots. They found that 30%–40% of automatic texts on factual topics deceive ordinary internet users (and 15%–25% deceive even experts), whereas this percentage goes up to 60% for non-factual (entertainment, adult) topics (30% with experts). They also find that texts that are disliked by the crowd have a higher deception rate (from 10% to 15% higher versus texts that are liked or rated as neutral) for both ordinary users and experts. These findings indicate that anti-

[26] Reddit users can rate comments by assigning likes and dislikes whose sum produces a comment score. The authors categorize Reddit comments as positive (positive score), neutral (zero score), and negative (negative score). They also randomly assign scores to artificially generated texts.

democratic computational propaganda in democratic countries has the potential to be harder to detect due to perceptional biases in both the general public and the expert community to view disagreement with the dominant viewpoint as a sign of human activity.

Technical Discussion: Can We Find Bots?

In order to study bot activity, of course, there is the prior, non-trivial, task of actually being able to find and identify bots. In this section, we briefly summarize these methods; readers should note that we include some highly technical language here for those who are interested.

Scholars have developed a number of bot-detection systems (Ferrara et al. 2016). Most of these are designed for use with Twitter data, which is in part a function of the importance of Twitter for political communication, but in the equal part is an artifact of its easy-to-use API (Application Program Interface), which makes it possible to access tweets, their metadata, and their related network data (although the latter is a computationally intensive process without access to the Twitter firehose).

As this research started only recently, there is a large degree of variety in the methods employed. Both supervised and unsupervised learning is utilized. Text and non-text features are used in detection. Some systems rely primarily on network data, others try to use linguistic tools to analyze the contents of the tweets, and still others capitalize on account behavior and other non-text features. Moreover, some systems are trying to do real-time classification, while others create algorithms for classifying data that was previously stored and is ready for processing.

Examples of unsupervised bot detection algorithms include DNA research-inspired sequence methods (Cresci et al. 2016) and applications of Dynamic Time Warping distance to identify coordination in accounts' activities (Chavoshi et al. 2016). Supervised methods build upon a plethora of algorithms including penalized generalized linear models, classification trees, support vector machines, boosting methods, etc. (Stukal et al. 2017; Ratkiewicz et al. 2011a; Chu et al. 2012; Oentaryo et al. 2016). Examples of real-time systems are Botometer (formerly, BotOrNot, Davis et al. 2016), Hoaxy (Shao et al. 2016), TwitterTrails (Finn et al. 2014), RumorLens (Resnick et al. 2014), TweetCred (Castillo et al. 2011), and Truthy (Ratkiewicz et al. 2011b). There are also systems of more generic use developed in the field of computational journalism, including FactWatcher (Hassan et al. 2014).

Most of these systems have been applied to one or a few datasets and, in most cases, they were purposefully designed for those datasets. Expectedly they usually demonstrate relatively high precision and recall. Assessing how these systems perform on new data and in answering different kinds of research questions (especially going beyond bot detection and instead focusing on analyses of bots' interactions with humans, impact, and strategy), and creating synthetic methods with wider applicability, are the next logical steps in this line of research.

However, such future systems will face a number of important challenges. The first, both

from a theoretical and empirical standpoint, is the lifespan of any given method, given that bots also change and grow in sophistication. Stukal et al. (2017) in a preliminary analysis demonstrate that their classifier loses about 20% of its precision if a training set from one year is applied to data from the following year, even in the same country. However, more systematic research is needed to shed light on this issue.

Secondly, despite the relative ease of accessing data, bot detection on Twitter is not a trivial computational task, partly because the automation of the Twitter feed makes it harder to separate legitimate human users who choose to use some automatic functionality from a fake account operated by a bot (Chu et al. 2012; Radziwill & Benton 2016). There is a consensus in the literature that this problem will only deepen with hybridization between humans and algorithms (Grimme et al. 2017).

Finally, computational propaganda and the dynamics of misinformation on social media so far have mostly been studied with respect to Twitter because of its easy-to-use API. In order for similar research to spread, for example, to Facebook, the research community will likely need a more open Facebook API. Indeed, changes in privacy policies implemented by social media platforms may be needed to give scholars enough information to detect and address bots in real time.

Platforms' Vulnerability to Disinformation

Before closing, we want to highlight two issues that make social media platforms particularly vulnerably to disinformation campaigns.

The first is a business model focused on ad revenue. History suggests that in absence of specific government regulations concerning who can and cannot advertise, companies will chase the revenue. For example, ahead of the 2016 U.S. presidential elections, Twitter reportedly offered the Russian state-supported media network RT 15% of its advertising for $3 million,[27] and Facebook demonstrated little interest in screening advertisers on its platform, allowing ads to be paid for in Russian rubles.[28] Anecdotal evidence suggests that platforms located in Russia took a very different approach to monitoring political advertisements.[29]

---

[27] Kantrowitz, Alex. November 1, 2017. "Twitter Offered Russian Television Network RT 15% of its Total Share of US Elections Advertising." *BuzzFeed*. https://www.buzzfeed.com/alexkantrowitz/twitter-offered-rt-15-of-its-total-share-of-us-elections.

[28] Smith, David. October 31, 2017. "Angry Al Franken Hammers Facebook Lawyer at Hearing over Russian Ads." *The Guardian*. http://www.theguardian.com/us-news/2017/oct/31/facebook-russia-ads-senate-hearing-al-franken.

[29] One of the authors of this review had an experience of trying to run political ads on Facebook, Google, Yandex ("Russian Google"), VK, and Odnoklassniki (the last two could be both called "Russian Facebooks") in 2015. Notably, the attempt to place ads was part of a political science experiment in Belarus—a country that is "foreign" for both American and Russian companies (Belarus is a close ally of Russia, but Russian companies appeared to have the same set of rules for all foreign countries). Facebook and Google ran all the requested ads, including featuring pictures of country political leaders and headlines critical of them. Yandex also ran the ads, but some pictures, including one with political prisoners on it, were rejected. Odnoklassniki

By contrast, registration requirements depend more on the business and content model adopted by the platform, and thus are not an "inherent" vulnerability. While most social media platforms adopt special registration procedures (CAPTCHA, email credentials, IP address requirements) designed to prevent the creation of fake accounts, they differ in the amount of information and kind of verification they seek from users. Early on, Facebook incentivized users to reveal their real name, location, educational background, and other biographical information, and instituted strict verification procedures. Twitter, on the other hand, encouraged tweeting under a nickname and required only minimal information about its users. These differences are reflected in the cost of followers that are available for purchase on the online black markets. Paquet-Clouston et al. (2017) claim that the average price for 1,000 followers is $15 on Twitter, $16 on Instagram, $34 on Facebook, and $49 on YouTube.[30] (Earlier research by Thomas et al. [2013] found that the price for 1,000 Twitter accounts ranged between $20 and $100, and that merchants have raised hundreds of thousands of dollars selling them.)

To satisfy the demand for accounts and followers, online merchants use both technical and non-technical means of getting around registration barriers and requirements. Thomas et al. (2013) show that merchants on the black market own or rent access to thousands of different hosts to get around the restrictions on IP addresses. Additionally, Paquet-Clouston et al. (2017) show that creators of bot accounts can efficiently pass phone verification when creating fake accounts using Voice Over Internet Protocol (VoIP) services. Sometimes accounts are bought and/or used much later than when they were created. A non-trivial number of bots tweeting about Russian politics in 2014–15, which were identified by Stukal et al. (2017), were created in the early 2010s and even late 2000s. Since accounts of real people could carry higher trustworthiness and reach a wider audience, merchants often seek to rent or buy real people's Twitter and Facebook accounts. The Russian Embassy in London has even created "Russian Diplomatic Online Club," whose members sign up for automatically retweeting ambassadors' tweets.[31]

A second important factor for platforms' vulnerability to disinformation campaigns is their optimization algorithms. Optimized for engagement (number of comments, shares, likes, etc.), they often help in spreading disinformation packaged in emotional news stories with sensational headlines. In a widely publicized analysis, Buzzfeed found that in the last three months ahead of 2016 U.S. presidential election, 20 top-performing fake news stories generated 8.7 million shares, reactions, and comments, while 20 top-performing stories from reputable news outlets generated a total of only 7.3 million shares, reactions, and

---

refused to run any ads as they completely prohibit political advertisement. VK also rejected all but the most tepid ads (featuring economic news, not political slogans) and explicitly stated that anything "critical of politicians, their political activity, or governance" is strictly prohibited. It also has a specific policy that photographs of famous people could be used only if these people provide written consent. It applies to public officials, including political leaders of the country, effectively barring anyone but their own campaigns from using their photographs in ads. They also bar using country flags in ads.

[30] Subscribers in the case of YouTube.

[31] Sullivan, Ben. March 15, 2017. "The Russian Embassy Is Asking People to Become Twitter Bots." *Motherboard*. https://motherboard.vice.com/en_us/article/jpnvax/the-russian-embassy-is-asking-people-to-become-twitter-bots.

comments (Silverman 2016). In a similar analysis ahead of the 2017 German parliamentary elections, Buzzfeed found that seven out of the ten most shared articles about Angela Merkel on Facebook were false (Schmehl & Lytvynenko 2017).

Another tactic is to produce a catchy fake image (typically, a photo) that would be actively reposted in social media. Gupta et al. (2013) studied more than ten thousand tweets with URLs to fake images and found that 86% of them were retweets, whereas only about 14% of users involved in the dissemination of misinformation posted the original tweets. Additionally, the detection of fake images on Twitter and Facebook is complicated by the fact that both platforms remove Exif[32] metadata (date, time, and location of image production, device model, and copyright information) from the posted images, whereas this metadata is the basis of most of the forensic techniques in the cases of digital images (Boididou et al. 2017; Huckle & White 2017).

While many, including the online platforms themselves, acknowledge the problem of their algorithms amplifying fake news, no consensus has emerged yet on an optimization criterion other than engagement.[33] In fact, the most widely discussed proposal—to flag suspected fake news stories or, even more radically, filter them out—alters the universe of stories user can potentially engage with, rather than creating a different metric for engagement. Even this strategy, however, faces a number of very significant challenges.

Firstly, the quality of verification, whether manual or automated, could suffer from both unintentional errors and systematic bias, removing legitimate content and letting slip deceptive content[34] . Manual verification by editors and professional fact-checkers could easily lead to perceived or real censorship. In Sina Weibo, users are encouraged to report suspicious news, and a committee composed of reputable users is supposed to judge the case (Jin et al. 2014; Jin et al. 2016), creating an obvious avenue for censorship. Facebook attempted to curate newsfeed of its users, which quickly lead to the accusations of anti-conservative bias.[35]

However, crowdsourcing or automated approaches have not been able to boast a better track record so far. Facebook has been accused of blocking legitimate accounts after an organized mob reports it as offensive for political reasons.[36] Similarly, the methodology

---

[32] "Exchangeable image file format" (Exif): "standard that specifies the formats for images, sound, and ancillary tags used by digital cameras (including smartphones), scanners and other systems handling image and sound files recorded by digital cameras." (https://en.wikipedia.org/wiki/Exif)

[33] Oremus, Will. January 3, 2016. "Who Controls Your Facebook Feed." *Slate*. http://www.slate.com/articles/technology/cover_story/2016/01/how_facebook_s_news_feed_algorithm_works.html.

[34] Levin, Sam. May 16, 2017. "Facebook Promised to Tackle Fake News. But the Evidence Shows It's Not Working." *The Guardian*. http://www.theguardian.com/technology/2017/may/16/facebook-fake-news-tools-not-working.

[35] Nunez, Michael. May 9, 2016. "Former Facebook Workers: We Routinely Suppressed Conservative News." *Gizmodo*. https://gizmodo.com/former-facebook-workers-we-routinely-suppressed-conser-1775461006.

[36] Among many others, the organizer of a feminist flashmob in Ukrainian Twitter was temporarily blocked: http://gordonua.com/news/society/facebook-zablokiroval-akkaunt-organizatora-fleshmoba-yaneboyusskazati-melnichenko-140855.html.

employed to detect deception on Sina Weibo was based on applying topic-viewpoint modeling (Trabelsi & Zaiane 2014) and treated as suspicious news that produces conflicting reactions. Although this approach has potential in detecting deception about some facts, it has serious drawbacks when applied to politics, where we expect the discussion of matters of political debate to feature multiple conflicting viewpoints. An additional layer of complexity comes with applying traditional deception detection techniques to social. Previous research has shown that one of the features that distinguish deceptive texts from truthful ones is its verbosity (Zhou & Zhang 2008). However, the limit on the length of tweets set by Twitter makes them a very special form of writing that is hard to analyze with deception detection algorithms (for a review, see Rubin 2017).

However, even if verification is done and delivered to potential readers, research shows they are not bound to accept the judgment and reject fake news. Indeed, models developed so far predict that even if users want to share only truthful news, fake news articles can "attain 'truthful news status' and […] propagate in perpetuity" (Papanastasiou 2017). Empirical evidence, both at the macro- and micro-level, is mixed, but includes cases when attaching flags and warnings had minimal (Pennycook & Rand 2017) to no effect (Vargo et al. 2017), with heterogeneous effects across demographic and partisan groups, including those for whom the effect was the opposite—i.e. the warning backfired. In addition, Pennycook and Rand (2017) document a potentially worrisome "implied truth effect," where articles without warnings were "seen as more accurate than in the control,"

Finally, as disinformation campaigns that rely on trolls and, especially, on bots require centralized coordination of a very large number of accounts, platforms that allow—for perfectly legitimate advertising purposes—integration with third-party social media management dashboards end up significantly reducing the costs and increasing efficiency of botnets and troll factories. For example, Stukal et al. (2017) find that using *dlvr.it* (previously, *twitterfeed*) was a very strong predictor of bots in a large dataset of tweets about Russian politics (see also Radziwill & Benton 2016).

E. Online Content and Political Polarization[37]

Executive Summary

This review is organized around six thematic areas: partisan cues, group cues, emotional cues, exposure and recency, virality, and audiovisual content. The main findings emerging from this review are as follows. (1) The prevailing consensus in political science is that elite behavior, rather than communication, is driving political polarization. That being said, messages that emphasize inter-party conflict reinforce polarization, while messages that stress intra-party conflict have the potential to reduce it. Partisan cues can also encourage partisans to accept and propagate inaccurate information. (2) Messages priming group cues and stereotypes can facilitate acceptance of inaccurate information about the out-group. (3) Emotions are important: Anger makes people less likely to distrust inaccurate information that supports their views, and more likely to distribute it; anxiety can have the opposite effect, prompting individuals to pursue accuracy rather than directional goals. (4) The volume and recency of disinformation matter: People are more likely to be affected by inaccurate information if they see more and more recent messages reporting facts, irrespective of whether they are true. (5) Viral mass-scale diffusion of messages is relatively rare. Information achieving mass spread usually relies on central broadcasters in a network and/or amplification by the mass media. Communities of belief, such as conspiracy theorists, are important in generating the kind of sustained attention that is needed for false information to travel. Content that is highly controversial is more likely to be shared by social media users. (6) There is reason to believe that audiovisual messages can be both more persuasive and more easily spread than textual messages, but we do not know nearly enough about these dynamics—most research to date has focused on textual rather than visual and audiovisual misinformation.

Partisan Cues

Partisan cues in news coverage of politics have been found to contribute to polarization by increasing the salience of partisan attitudes. Partisan media has been studied extensively in this regard. Levendusky (2013) argues that by presenting politics as a struggle between irreconcilably opposed parties, partisan media make audiences' partisan identities more salient, thus contributing to both cognitive and affective polarization (see also Stroud 2011). Relatedly, Garrett et al. (2016) show that exposure to ideologically slanted websites is positively associated with holding inaccurate beliefs on politically relevant issues *even if individuals are aware of the correct facts contradicting such beliefs*.[38]

---

[37] Review prepared by Cristian Vaccari, Reader in Political Communication, Loughborough University, United Kingdom, and Associate Professor of Political Science, University of Bologna, Italy.

[38] Although see Review F (below), as well https://slate.com/health-and-science/2018/01/weve-been-told-were-living-in-a-post-truth-age-dont-believe-it.html, for discussions about new evidence (and reanalysis of old evidence) suggesting that it may be easier to change people's opinions than previously believed.

Mainstream news coverage, however, can have similar polarizing effects. In a review of research on this topic, Arcenaux and Johnson (2015) argue that news stories generally report where parties stand on issues, and if party elites' issue stances are polarized, news coverage is bound to reflect this, regardless of whether the news outlet producing it is partisan or mainstream. To the extent that voters take cues from elites, "party elites may bear more of the responsibility for the polarized state of the country. News media, including mainstream and partisan outlets, are megaphones more than motivators of partisan polarization" (Arcenaux & Johnson 2015, pp. 322-3). In this regard, the mainstream news practice of giving voice to both sides of a controversy may make it even clearer to viewers that elites are starkly divided across party lines (see also Prior 2013). Relatedly, Garrett and colleagues find that engaging with both pro- and counter-attitudinal websites was positively associated with in-group favorability more strongly than exclusive exposure to pro-attitudinal websites, while engagement with counter-attitudinal websites was negatively associated with in-group favorability (Garrett et al. 2014). In other words, users who hear both their side and the other side tend to be even more convinced of the validity of their own views than those who only get news from sources confirming their opinions.

In politicized environments, different message cues can help prime directional goals (achieving attitude consistency even in the face of ambiguous or attitude-disconfirming information) or accuracy goals (developing beliefs based on information one believes to be true).

In an experiment, Druckman et al. (2013) manipulated the kinds of arguments subjects were exposed to (distinguishing between weak and strong arguments) as well as information on the level of polarization on the issue among Republicans and Democrats in Congress. When subjects were not told that elites were polarized, they changed their attitudes toward the stronger argument they had been shown. When subjects were told that elites were moderately polarized on an issue, they still followed the stronger arguments when they were exposed to them, but when they saw weaker arguments they tended to revert to their parties' position. Finally, when subjects were told that elites were deeply divided on the issue, they tended to change their attitudes consistently with what they were told their party argued, irrespective of the strength of the argument—i.e., even if they had been exposed to a stronger argument from the out-party. Relatedly, Brulle et al. (2012) find that U.S. public opinion on the threat of climate change was moved more by elite cues—in particular, Congressional Republicans' opposition to climate change bills—than by media coverage, which by and large mirrored those cues while also presenting Democrats' position.

Partisan cues also play an important role in voters' likelihood of believing in rumors. Weeks and Garrett (2014) find that "exposure to rumors about the candidates is positively related to belief [in said rumors] for members of both parties, but the relationship is significantly stronger when the rumor is attitude-consistent" (p. 409). Partisanship, however, is also important in the opposite scenario, when the source and the content of the message contradict voters' expectations. In those infrequent situations, messages are more credible than in more ordinary situations when partisan elites behave as voters normally

expect them to. Thus, Berinsky (2017) demonstrates that rumors are more effectively corrected by "unlikely sources"—that is, people who argue against their personal and political interests—than by sources who can be expected to be opposed to the content and political implications of the rumor. Republican politicians' corrections to the false "death panel" rumor in the debate about the Affordable Care Act were more effective than non-partisan and Democratic politicians' corrections. Similarly, Baum and Groeling (2009) find that news coverage of "costly" internal party disputes and elites' positions that run counter their parties' interests (as when partisan elites criticize their fellow partisans or praise members of the other party) is more credible than "cheap talk" that shows elites toeing the party line. Moreover, such "trespassing" messages are more credible when they appear on news outlets that normally support the opposite positions.

Some message cues can prime audiences to resist politicization of scientific messages and respond to factually accurate information even when it contradicts their partisan preferences. In a survey experiment, Bolsen and Druckman (2015) found that subjects tend to disbelieve scientific evidence when it is presented as subject to partisan disputes, but when subjects are warned that scientific consensus is overwhelming, even in the face of partisan disputes, accuracy goals prevail over directional ones.

Uncivil messages have been found to lower perceptions of the legitimacy of the opposition's, but not one's own party's, arguments, thus augmenting affective polarization. In an experimental study of television talk shows featuring uncivil discussions between politicians and commentators, Mutz (2007) found that uncivil discourse led to viewers' acquiring greater awareness of both their own and the opposition party's positions, but also worsened subjects' evaluations of the opposition and perceptions of the legitimacy of their arguments. To the extent that partisan and misinforming online discourse is often uncivil itself, or accompanied by uncivil comments by other social media users, it is conceivable that similar effects may arise online, even though research is needed to verify whether Mutz's findings from television extend to digital media.

In sum, messages that emphasize partisan divides can increase polarization, regardless of the partisanship of the source and the audience watching it. By contrast, messages that emphasize intra-party disagreement can reduce polarization, and this may be especially the case when the sources of these messages normally take contrary stands. Finally, messages that warn audiences that, in spite of political divisions, scientific consensus is widespread have the potential to induce accuracy goals.

Group Cues

Negative attitudes toward groups are an important component of polarization, in both its cognitive and affective dimensions. Negative perceptions of certain groups may also enhance belief in false information about those groups. As argued by Kosloff and colleagues (2010), "When persons are viewed as distinctly different, negative labeling can be accomplished smoothly because there is little harm in attributing all manner of bad characteristics to 'them'" (p. 384). Simply reminding subjects of the groups they belong to might enhance their likelihood of accepting false information about out-group members,

even if the identity of such out-group has not been made explicit. Thus, Kosloff et al. (2010) find that making age salient increased undecided voters' likelihood to believe the smear that John McCain was senile during the 2008 presidential campaign, while making race salient increased undecideds' propensity to believe the smear that Barack Obama was Muslim. Priming these group cues also increased partisans' likelihood to believe those smears, but only along party lines, with Republicans more likely to believe Obama was Muslim and Democrats that McCain was senile. Moreover, priming race also increased both Republicans' and undecideds' propensity to believe Obama was a socialist—a kind of disinformation not directly related to the (racial) group cue primed by researchers. This suggests that out-group cues may elicit negative political beliefs and facilitate manipulation around seemingly unrelated issues.

One way to counteract polarization and belief in inaccurate information that primes negative group attributions is to develop messages that can improve negative attitudes toward out-groups. In a series of experiments, Wojcieszak and her collaborators tested how different message features can improve respondents' evaluations of groups they dislike. In a U.S.-based study, Wojcieszak and Kim (2016) show that counter-attitudinal messages based on narratives emphasizing personal stories and experiences are more likely to be accepted by subjects than messages based on numbers (both generalizable statistics and specific data points). Narrative messages are more effective when subjects are encouraged to empathize with the out-group members, whereas messages based on numbers are more likely to provoke attitude change when subjects are encouraged to evaluate the issues objectively in a detached way. According to Galinsky and Moskowitz (2000), when individuals are prompted to take the perspective of out-group members, they become less likely to resort to stereotypes to describe them, and tend to be less biased in their views of in-group and out-group members. This may explain why narrative messages may improve attitudes toward out-group members by encouraging subjects to take the perspective of those people. In a study on Muslim immigrants to the Netherlands, Wojcieszak et al. (2017a) find that Dutch-born, second-generation migrants are more likely to change their minds on gender equality, sexual minority rights, and secularism in public life when they are exposed to narrative messages, while first-generation migrants are more likely to respond to numbers-based messages. They interpret these differences as the result of different cultural orientations, as more Westernized second-generation immigrants are more likely to espouse individual-centered narratives, while first-generation immigrants are more comfortable adopting the kind of holistic thinking that statistical evidence encourages.

Citizens also encounter important cues on out-groups via the mass media. Because people tend to gravitate around other people that resemble them socially, ethnically, and culturally, the mass media often provide the only source of information on more distant groups (Mutz & Goldman 2010). Wojcieszak and Azrout (2016) find that Dutch voters who were exposed to media coverage of Muslim and Polish immigrants developed more positive views of them—measured as social distance and perception of threat from the out-group. The effect was stronger when media coverage was positive, and worked above and beyond whether subjects also experienced face-to-face contact with the out-group. They also noted that mediated contact with immigrants in the context of crime stories increased social

distance and perceived threat, while seeing migrants featured in news about culture decreased those negative attitudes (p. 1053).

Group cues are thus important in eliciting partisan directional goals, which also leave subjects more exposed to the threat of disinformation by making attitude-congruent rumors more believable. Strategies are available to reduce negative attributions of out-groups, which may be a more viable route to reducing the negative impact of group cues than attempting to suppress group cues from political communication.

Emotional Cues

The emotions felt by audiences while they are exposed to a message play an important role in enhancing the message's credibility. In an experimental study, Weeks (2015) finds that "Anger encourages partisan, motivated evaluation of uncorrected misinformation that results in beliefs consistent with the supported political party, while anxiety at times promotes initial beliefs based less on partisanship and more on the information environment" (p. 699). Thus, messages eliciting anger are more likely to increase the salience of partisan cues and activate directional goals, while messages eliciting anxiety are more likely to activate accuracy goals where ascertaining the truth matters more than reaffirming one's partisan identity. Emotional arousal has also been found to increase social diffusion of information (Berger 2011), which suggests that emotionally charged messages have a higher probability to of becoming viral. Based on experiments, Heath et al. (2001) find that individuals are more likely to pass along urban legends that evoke feelings of disgust. Hasell and Weeks (2016) analyze panel survey data and find that respondents who used pro-attitudinal partisan news reported higher levels of anger toward the opposing party and that such anger was positively associated with subsequently sharing news on social media.

Emotions also contribute to the indirect spread of information via social media. Bail (2016) tracked the numbers of users who saw messages posted on Facebook by advocacy organizations around autism spectrum disorders. He found that when these posts had emotional features, they evoked emotional comments from those who followed the organizations on Facebook, and that these emotional comments, in turn, attracted "viral views," i.e., views of the message by other users who are friends or followers of the commenters, but not of the advocacy organizations. This was true for both positive and negative emotions. This is an important mechanism because it entails messages' ability to spread beyond self-selection by social media users. Relatedly, Brady et al. (2017) find that including "moral-emotional" words in tweets on three political polarizing issues (gun control, same-sex marriage, and climate change) made these messages significantly more likely to be shared on Twitter.

While public discourse around emotions in politics tends to equate emotions with lack of thinking and generally negative outcomes, political science research suggests a more nuanced picture where some emotional states (anxiety) yield cognitive benefits and may lead social media users to be more considerate about what they view and share, while other emotional states (anger) tend to enhance directional goals and facilitate polarization

and the spread of misinformation. Whether and how specific messages elicit these different emotional states is an area ripe for future research.

Exposure and Recency

Some research suggests that simple exposure to false information can make it more credible. This is because repetition increases processing fluency, which in turn is used as a heuristic to infer accuracy. Berinsky (2017) finds that fluency is a powerful factor in increasing recall and belief in rumors, and that some corrections, by increasing fluency, may enhance rather than reduce false beliefs. Pennycook et al. (2017) show that experimental subjects who saw false news headlines more than once were significantly more likely to treat them as accurate than those who saw them for the first time. These effects persisted even if subjects had received a preliminary warning that the news they were exposed to were disputed, and remained visible in a follow-up study one week later.

Exposure to news coverage about a topic may also polarize audiences irrespective of the tone. Wojcieszak et al. (2017b) find that exposure to news coverage on the European Union polarized citizens in the Netherlands who held the most extreme, both pro- and anti-E.U., positions. The effect of news coverage was stronger on the diffuse dimensions of E.U. attitudes (i.e., identity and negative affections) than the specific dimensions (i.e., utility and performance).

Recency of messages can also play a role. In another experiment based on a student sample, Westerman et al. (2014) found that subjects exposed to different Twitter feeds were more likely to trust those that were more recently updated. Recency prompted subjects to engage in higher levels of cognitive elaboration of the messages, which in turn was positively associated with the credibility attributed to the source.

These results suggest that frequent repetition of false or polarizing information can achieve greater effects, all else being equal, by both increasing fluency among those that encounter messages multiple times and looking more up-to-date to those that encounter them for the first time.

Virality

As a premise, it is important to realize that virality, understood as growth in the diffusion of a message through person-to-person contacts similar to the spread of a disease, is not the most common mechanism by which information spreads in online networks. Goel and colleagues analyzed a billion diffusion events on Twitter and found that the main reason messages spread is that they are shared by "broadcasters," or users who have large audiences, while the "viral" model, where messages achieve mass diffusion via large numbers of individual peer-to-peer transmissions, is less common (Goel et al. 2015). Jiang et al. (2014) found one of the characteristics that predict the popularity of an online video is the popularity of the user who posted it, rather than the time when it was posted—many viral videos have duplicates posted by different users, and the most popular video is not necessarily the first that was uploaded if its uploader was not particularly popular.

Research by Rojecki and Meraz (2016) finds that the web is not always sufficient to propagate misinformation at mass scale, but it can be aided by the mass media, so online sources can have an important role in seeding false stories that go viral only after they have been covered by the mass media.

Virality is easier to achieve at the beginning of a high-profile event or crisis, when many people are paying attention, but trusted authorities (police, scientists, journalists) have not yet provided an authoritative narrative to explain the situation and recommend specific courses of action. In an information and knowledge vacuum, rumors quickly fill the void. One widely accepted definition of rumors emphasizes the role of event-related uncertainty for their spread: "unverified and instrumentally relevant information statements in circulation that arise in contexts of ambiguity, danger, or potential threat and that function to help people make sense and manage risk" (DiFonzo & Bordia 2007; cited in Silverman 2015). This is why misinformation spreads easily in the early stages of disease epidemics, when people feel the need for explanations of new and unknown phenomena.

Both true and false information is propagated online via informational cascades whereby individuals share messages in a way that makes diffusion grow exponentially until it reaches a peak. Cascades on social media normally involve groups of like-minded users, or at least users who gravitate around the same social media profiles, but can also involve cross-cutting ties between individuals who are only loosely connected to each other (Colleoni et al. 2014). Del Vicario and colleagues (2016) find that news about scientific discoveries and conspiracy theories follow similar paths, whereby diffusion peaks relatively early (within the first two hours after the information was originally seeded) and then declines rapidly. However, "Science news is usually assimilated, i.e., it reaches a higher level of diffusion quickly, and a longer lifetime does not correspond to a higher level of interest. Conversely, conspiracy rumors are assimilated more slowly and show a positive relation between lifetime and size" (p. 556). In other words, conspiracy theories require sustained attention and distribution by their supporters to reach critical mass, while scientific information does not.

Understanding the groups of social media users that can generate this kind of sustained attention and diffusion thus becomes important. Believers in and propagators of misinformation tend to focus on specific topics. Bessi et al. (2015) find that Italian supporters of conspiracy theories on Facebook concentrate their social media activities on four thematic areas—environment, diet, health, and geopolitics—and that most members of conspiracy communities engage at similar levels with posts related to all four topics. Better understanding the topics around which conspiracy theorists congregate in different countries (and regions) may help predict which kinds of content are more likely to go viral on social media, as messages focusing on conspiracy theorists' preferred topics can count on a willing army of supporters and spreaders.

Research on the factors that lead polarizing or disinformation messages to go viral is still lacking, but we can rely on some experimental research on the factors that facilitate the sharing of a message irrespective of its truthfulness or polarizing nature. Evidence suggests messages whose content stands out from the normal flow of information are more likely to

be circulated. Rudat et al. (2014) find that Twitter users are more likely to share messages that contain high informational value factors like controversy, relevance, or unexpectedness—values that also increase the likelihood that a story is covered by news organizations. Content that is outrageous and counterintuitive is thus more likely to be shared, if believed. In a survey on Twitter users who shared tabloid news during the 2017 U.K. general election, Chadwick et al. (2017) found that users who were motivated by the desire to debate—to find out other people's opinions and provoke discussions—and those aiming to provoke others—by entertaining, pleasing, or upsetting them—were significantly more likely to admit sharing news that was inaccurate or exaggerated. A survey of Singapore university students by Chen et al. (2015) similarly reveals that catchiness and the ability to spark conversations were key motivators for news sharing on social media.

Social endorsements are also important. Metzger et al.(2010) find that internet users rely heavily on the "endorsement heuristic," whereby "people are inclined to perceive information and sources as credible if others do so also, without much scrutiny of the site content or source itself" (p. 427). Li and Sakamoto (2014) find that exposing people to information about how likely other users are to share a message positively influences subjects' intention to share that message themselves. Importantly, subjects followed these endorsement cues to the same extent irrespective of whether they perceived the statement they were presented to be true, debatable, or false. By contrast, when subjects were not exposed to endorsement cues, they were less likely to share statements they thought were false. This suggests that social endorsement cues (such as numbers of likes and retweets) may enhance the credibility of false information even when individuals are unsure about its veracity.

Thus, all else being equal, inaccurate or polarizing content can be expected to be more likely to spread if users believe many other people are sharing and endorsing it. The role of social media bots, as well as committed networks of extremist activists and conspiracy theorists, in propping up the numbers of shares and likes of unverified content must thus be thoroughly investigated, as these forms of "digital ballot stuffing" may activate the endorsement heuristic and increase the likelihood that unverified information is both believed and shared by other users.

Audiovisual Content

Most of the research on the diffusion and effects of polarizing and misinforming messages focuses on the *textual* rather than the *visual and audiovisual* component of these messages. Yet substantial amounts of social media content nowadays are visual and audiovisual, and visual content is more likely to be shared than textual content. According to industry data, infographics are liked and shared on social media three times more than any other type of content; tweets with images receive 150% more retweets than tweets without images; articles with an image once every 75–100 words receive double the social media shares as articles with fewer images; and Facebook posts with images generate 2.3 times more

engagement than those without images.[39] Goel et al. (2015) find that cascades involving videos and pictures tend to achieve higher popularity than cascades of users sharing news and petitions.

We have known for a long time that human beings recognize and remember pictures more easily than words. Pictures are richer in stimuli than textual and verbal content, and they are processed more effectively by the brain (Stenberg 2006). Sundar (2008) argues that users process audiovisual content based on the "realism heuristic," as they assume that audiovisual content has a higher resemblance to the real world than textual and verbal content. Images, however, can also be easily doctored or presented out of context because viewers believe them to speak for themselves. Some studies offer anecdotal evidence of the role of visuals in the spread of misinformation. For instance, Zubiaga and Ji (2014) find that users had difficulty detecting the authenticity of doctored photos shared by social media users during Hurricane Sandy in 2012. Images are often taken out of context on social media. However, we know very little about the dynamics of the spread of visual misinformation besides anecdotal evidence and case study research.

Even more problematic in this regard is the rapid development of technologies that can synthesize audiovisual clips of human speech that closely resembles real speech based on relatively small training sets of original video (Thies et al. 2016).[40] If human beings cannot distinguish between original and synthesized audiovisual content, and if audiovisual content is more likely to be shared, watched, and remembered than other types of content, the diffusion of these technologies may have a much bigger potential to mislead users than textual content.

[39] See https://blog.hubspot.com/marketing/visual-content-marketing-strategy (accessed December 7, 2017).

[40] See http://www.washington.edu/news/2017/07/11/lip-syncing-obama-new-tools-turn-audio-clips-into-realistic-video/

F. How Misinformation and Polarization Affect American Democracy[41]

Executive Summary

Partisan polarization has increased dramatically at the mass and elite level since the mid-20th century in the United States, producing important and largely unanticipated challenges for American democracy. The norms of political institutions are being deeply strained by intense elite partisanship. At the mass level, greater partisan divisions in social identity are generating intense hostility toward opposition partisans that encourages extreme tactics and undermines compromise and civility. These developments have seemingly increased the political system's vulnerability to partisan misinformation, which is often promoted by polarized elites to sympathetic partisan audiences. Widespread usage of social media and distrust of the media threaten to accelerate these trends.

Effect of Polarization on Democratic Performance

I first consider the effect of increased polarization on democratic performance in the U.S.[42] Before doing so, however, it is essential to put the recent increase in polarization into a broader context. Most discussions of the topic start with—and bemoan—the increase in polarization since the mid-20th century, failing to recognize that the low polarization observed in the postwar period was a historical anomaly rather than a norm that has been disrupted. Party polarization in the mid-20th century plunged to unprecedented levels; for instance, differences between the parties in Congressional voting patterns reached a historic low for the post-Civil War period (see, e.g., Bonica et al. 2013). This change was closely linked to the issue of race, which increasingly divided the parties internally even as one-party status of the Jim Crow South kept conservative Southern legislators in the Democratic Party. The result was that the parties of this period were broad and heterogeneous coalitions that appeared indistinct to many voters.

In response to the ideological overlap between the parties, the American Political Science Association (APSA) famously issued a report calling for *more* polarization (1950). Specifically, a committee convened by the association called for the parties to propose specific public policy programs that would offer clear choices to voters and to seek to implement those while in office. Voters would then in turn be able to hold parties responsible for their actions in office. To accomplish these goals, the APSA committee notes that the parties would need to develop greater unity and party loyalty than existed at the time and proposed a series of measures to do so.

[41] Review prepared by Brendan Nyhan, Professor of Government, Dartmouth College.

[42] Unless otherwise stated, "polarization" here refers to ideological polarization, or the distance between the parties on a left-right scale in terms of their policy preferences. Later in the section we discuss "affective polarization," or the extent to which supporters of one party dislike the other party.

In the years since the APSA report, the parties realigned on the issue of race and civil rights (Carmines & Stimson 1989), setting in motion a process by which moderate and conservative Democrats in the South were replaced with conservative Republicans. Ideologically motivated activists and party leaders helped capture and shape the parties during this period, increasing the magnitude of polarization and extending it to new issues (for a review, see Layman et al. 2006). As a result, the parties are further divided ideologically in Congress than ever (Bonica et al. 2013). Correspondingly, though many citizens still have relatively inconsistent preferences over issues, they are better sorted into parties based on ideology and support those parties more consistently across offices and levels of government (Fiorina & Abrams 2008; Abramowitz & Webster 2016). The parties are in turn perceived by the public as more clearly distinct—only 18% said there is no important difference between the parties in 2012 compared to 56% in 1966 (ANES Guide 2015).

These developments have in many ways satisfied the goals of the APSA committee. The parties have more coherent policy agendas than in the past and now provide relatively clear and distinct choices to voters. However, the increase in polarization observed in the U.S. has also had harmful effects on American democracy that the APSA task force and other observers did not fully anticipate.

First, as the parties have become more distinct in both their ideology and the demographic groups that support them, partisanship has become a potent social identity, driving feelings toward opposition party identifiers to new lows. This pattern of so-called "affective polarization" or "negative partisanship" has generated remarkable levels of hostility toward opposition party identifiers (Iyengar et al. 2012; Huddy et al. 2015; Iyengar & Westwood 2015; Mason 2015; Abramowitz & Webster 2016; Rogowski & Sutherland 2016). These negative feelings threaten to undermine norms of civility and mutual respect in political debate. The strongly negative affective reactions that opposition partisans now inspire create a constituency for the winner-take-all political tactics discussed below and undermine the incentives for elites to engage in civil discourse and policy compromise. When the opposition party is despised, it also limits accountability for own-party figures, whose failings and foibles can more easily be rationalized as better than the opposition.

Second, American political institutions and democratic norms have come under strain in this era of high partisanship. As the parties have grown more distinct and homogenous, they have exploited their use of procedural and agenda-setting powers to attempt to shift policy toward the median majority party member (Aldrich & Rohde 2000, Cox & McCubbins 2007). In response, minority party legislators have exploited the high number of veto points in the American system of government (most notably, the filibuster rule in the Senate) to block legislative action, making it difficult for presidents to enact legislation under divided government (e.g., Bond et al. 2015). Views on the merits of polarization and policy gridlock differ, but the intense form of elite party competition and polarization witnessed in recent decades has pernicious consequences, including the destructive use of scandal against opposition figures (Ginsberg & Shefter 1999) and reduced oversight over co-partisan administrations in the executive branch (Parker & Dull 2009, 2013). Previous norms limiting the range of acceptable tactics have also been breached in the use of

impeachment, high-stakes brinkmanship, and legislative hostage-taking (e.g., government shutdowns and debt ceiling standoffs), and extreme forms of gerrymandering and voter suppression in the states. This process of escalation has reached a new level under President Trump, who routinely violates norms pertaining to conduct by government officials that are vital to democratic governance (Nyhan 2017b). The extent to which those norms and institutions will constrain Trump remains to be seen.

Finally, the growth in polarization has seemingly supercharged political misinformation, leading to widespread partisan misperceptions and conspiracy theories that pollute public debate, distort public policy, and intensify polarization. (I discuss this point below and thus do not address it further here.)

Effect of Misinformation on Democratic Performance

Hundreds of books and articles catalogue the effects of polarization on American democracy, but the consequences of misinformation are less understood. Fears that voters lack the capacity to participate meaningfully, of course, go back to the beginnings of democracy, but until recently few studies have distinguished between being *un*informed and being *mis*informed (Kuklinski et al. 2000). However, we can identify a number of concerns about how misinformation could affect democratic decision making and policy processes that are normatively troubling.

First, misperceptions might distort the views of individual citizens. People frequently lack accurate information about politics and might hold different preferences or opinions if their views were more accurate. Counterfactual calculations suggest that ignorance distorts collective opinion from what it would be if people were better informed about politics, though these estimates rely on strong assumptions (Bartels 1996; Althaus 1998; Gilens 2001). Correspondingly, experiments providing people with accurate information about public policy issues do sometimes result in them expressing different opinions (e.g., Gilens 2001; Sides 2016). However, these conclusions need not hold; people often provide multiple rationales for their opinions and do not strictly base them on facts. Correspondingly, studies have shown that in some cases the provision of correct information has no effect on policy preferences on controversial issues (e.g., Kuklinski et al. 2000; Hopkins et al. N.d.) or on evaluations of high-profile partisan candidates (e.g., Nyhan et al. N.d.).

The effects of misperceptions at the individual level can aggregate into distortions in collective public opinion that likely affect policy and election outcomes. For instance, misperceptions about the state of the economy and the federal budget are widespread (e.g., Bartels 2002). The public may also misunderstand the state of wars overseas, especially early in a conflict (Baum & Groeling 2009). Though many misperceptions are self-generated (Thorson N.d.), these distortions may be created, encouraged, and/or exploited by political elites, who often seek to promote false or misleading claims in order to promote their preferred policies, win (re-)election, or avoid accountability for their performance in office (e.g., Fritz et al. 2004; Flynn et al. 2017).

Misperceptions can additionally distort the content of public policy debates. Salient examples are exaggerated perceptions about the generosity of U.S. federal welfare and foreign aid, and the number of immigrants in the country. A recent Kaiser Family Foundation poll found that Americans on average estimated that 31% of the federal budget goes to foreign aid; only three in 100 know the correct answer of less than 1% (DiJulio et al. 2016). A representative survey of Illinois residents conducted in the late 1990s found similarly that fewer than one in ten respondents knew that welfare spending amounts to less than 1% of the federal budget (Kuklinski et al. 2000). Americans similarly overestimate the size of the immigrant population (Hopkins et al. N.d.). Correcting these misperceptions may not immediately change people's opinions about these issues (Kuklinski et al. 2000, Hopkins et al. N.d.), but their existence and persistence likely affects the policy proposals offered by elected officials and the reactions they receive from the public.

How Polarization and Misinformation Interact

Partisan misinformation and conspiracy theories have seemingly increased in recent years in tandem with intensifying elite ideological polarization and widespread affective polarization at the mass level. Belief in these false and unsupported claims is frequently skewed by partisanship and ideology (see, e.g., Ramsay et al. 2010; Frankovic 2016, 2017), suggesting that our vulnerability to them is increased by directionally motivated reasoning—the tendency to selectively accept or reject information depending on its consistency with our prior beliefs and attitudes (Kunda 1990; Taber & Lodge 2006). Motivated reasoning can also undermine the effectiveness of corrective information, which sometimes fails to reduce misperceptions among vulnerable groups (e.g., contrast Nyhan & Reifler 2010 and Nyhan & Reifler N.d.; see Flynn et al. 2017 for a review). In the real world, disconfirming evidence may only temporarily decrease belief in misperceptions and can even increase them among vulnerable groups (Berinsky 2012; Schaffner & Roche 2016).

Many partisan misperceptions have become widespread and had significant effects on politics and public policy. In some cases, these may capitalize on other factors that increase vulnerability to misperceptions. President Obama, for instance, was plagued by myths that were grounded in perceptions of difference—first that he was a Muslim and later that he was not born in this country (Kosloff et al. 2010; Pasek et al. 2015). However, belief in the birther myth differed sharply by party, suggesting it was primarily a partisan myth facilitated by directionally motivated reasoning. The power of this form of reasoning is strong—belief in the myth among Republicans rebounded within weeks after the release of Obama's long-form birth certificate, a type of dispositive evidence that typically not available for other misperceptions, and continues to persist even now that Obama has left office (Berinsky 2012; Frankovic 2017). Similarly, many polarized policy debates are plagued by misinformation that hinders evidence-based debate. Two notable examples are the "death panel" myth, which affected both the debate over the Affordable Care Act and end-of-life policy more generally, and climate change, an issue in which many years of efforts to communicate the scientific consensus have failed to overcome polarizing elite messages that generate widespread disagreement by party and ideology (Nyhan 2010; McCright & Dunlap 2011).

These problems may become more severe if stronger directional preferences prompt people to engage in greater selective exposure to attitude-consistent information about politics (e.g., Stroud 2008; Hart et al. 2009; Iyengar & Hahn 2009; Iyengar et al. 2008). The prevalence of "echo chambers" in people's information diets is often exaggerated (e.g., Gentzkow & Shapiro 2011, Flaxman et al. 2016; Guess N.d.; see Guess et al. 2017 for a review) but social media and other online content formats and platforms may facilitate greater selective exposure (Bakshy et al. 2015), including to misleading information. Most notably, "fake news" was widely read and shared in the period before the 2016 presidential election (Silverman 2016; Allcott & Gentzkow 2017).[43] Behavioral data indicate visits to (overwhelmingly pro-Trump) fake news websites were heavily concentrated among a small subset of people with the most conservative information diets and were driven by exposure on Facebook (Guess et al. N.d.). Since the election, Facebook has undertaken a number of initiatives to limit the spread of fake news on the site, including a labeling initiative in partnership with fact checkers that appear to be at least somewhat effective (Pennycook & Rand 2017.; Pennycook et al. N.d.; Blair et al. N.d.), but it is unclear whether these approaches can effectively address the volume of dubious content on the platform without distorting the public's access to political information (Nyhan 2017a).

Another worrisome development is widespread distrust of the media, which has been fueled by the increasing flow of negative messages about the press from elites (Ladd 2011). These perceptions have become intensely polarized under President Trump, who regularly attacks the media in vitriolic terms and accuses it of fabricating stories. Trump supporters now report extremely low levels of trust in the media; large majorities believe the media fabricates stories, call the media an "enemy of the people," and say they believe it prevents leaders from doing their job well (Guess et al. N.d.b). Under these circumstances, it is extremely difficult for the press to effectively counter partisan misinformation.

Finally, perceptual distortions created by increased polarization and negative partisanship can create misperceptions about the parties that further increase political divisions in our society. One study finds that people overstate the extent of ideological polarization and report more moderate positions after being provided correct information about people's actual beliefs (Ahler 2014). In addition, negative partisanship may generate misleading stereotypes of opposition party identifiers. Partisans appear to hold distorted perceptions of the opposition party, whose motives they perceive to be very negative (Freeder N.d.); providing more positive information about motives reduces out-group hostility. Finally, partisans are especially prone to overstating the prevalence of party-stereotypical groups among the opposing party's supporters, such as LBGT individuals among Democrats and high-income individuals among Republicans (Ahler & Sood N.d.). Again, providing accurate information improves perceptions of the opposition party.

[43] Allcott and Gentzkow estimate that the average adult "saw and remembered" slightly more than one fake news story over the course of the 2016 election campaign (p. 213).

Section III: Looking Forward

A. Key Research Gaps

Executive Summary

In conjunction with preparing the literature reviews, each researcher was also requested to prepare a list of key research gaps in the area investigated. This section represents a synthesis of the suggestions across the different subject areas. It is presented in three parts: definitions, prevalence, and substantive research topics. Another way of thinking about this is that the first two sections (definitions and prevalence) represent preliminary groundwork that will better facilitate addressing all of the subsequent substantive research questions, reflecting a strong consensus across the reviews that there is important work to be done in this regard.

Key remaining research questions include:

1. What are the effects of exposure to information and disinformation on individual beliefs and behavior?

2. What are the cumulative effects of having accounts on multiple platforms, and how might such conclusions differ from what we've learned from studies of behavior on a single platform?

3. How does the spread of disinformation through images and video differ from the spread of disinformation through text?

4. How do the spread and the effect of disinformation differ across different countries?

5. Do the effects of exposure to disinformation and polarization vary across liberals and conservatives?

6. What are the likely effects of new laws and regulations intended to limit the spread of disinformation?

7. What are the strengths and weaknesses of different methods of bot detection and analysis?

8. What is the role of political elites in spreading disinformation online?

Definitions

One strong theme that comes out of all the reports is the fact that there is no real consensus across much of the academic literature on how to define many of the phenomena discussed in the report. Undoubtedly, research would benefit from a common set of definitions of the following topics:

- **Online conversations/interactions:** Strikingly, despite the pervasive belief that "online conversations" have gotten more antagonistic as a result of political polarization, we lack any real consensus as to what exactly is an online political conversation (see discussion in Review A). As part of addressing this topic, it would therefore be useful to have definitions for:
    - Online political conversations
    - Cross-partisan online conversations
    - Antagonistic or "uncivil" interactions
    - Echo chambers[44]

- **Disinformation**: Despite all the attention to disinformation, fake news, etc., we are still lacking common definitions for many of these terms (Born & Edgington 2017), which could include:
    - Disinformation (knowingly false information?)
    - Misinformation (unwittingly false information?)
    - Online propaganda (information intended to promote one party/candidate?)
    - Hyperpartisan news (news packaged to denigrate the other party?)
    - Fake news (false information produced to maximize clicks for profit?)
    - Clickbait (non-false information presented to maximize clicks for profit?)
    - Rumors (non-confirmed information?)
    - Conspiracy theories (false stories repeated over time with known contrast to receive wisdom, includes reference to fact that others are trying to suppress the truth?)

- **Media Classifiers**: Closely related to "online propaganda," we are seeing increasing instances of the use of the term "hyperpartisan media." It seems important then to have a clear set of definitions so different studies examining the effects of media actors have similar conceptions of the following categories:
    - Hyperpartisan media
    - Partisan media
    - Non-partisan media

[44] To date, not only is there no consensus on what level of selective exposure constitutes an "echo chamber," there is not even any consensus on what metric or summary statistic should be used to measure this selective exposure. One possibility is the overlap coefficient (OC), which characterizes the degree of overlap between two probability densities (e.g., liberal and conservative media diet distributions).

- **Online Actors:** Sanovich et al. (2018) propose a five-part categorization of Twitter accounts consisting of: official accounts (representing organizations); humans; bots (algorithmically controlled accounts); cyborgs (accounts with content produced by humans and bots); and spam (accounts that produce only advertising), which probably could be extended, with minor modifications, to other social media platforms as well. More generally, there seems to be a developing consensus in the literature as to what constitutes a "bot," less consensus on what constitutes a "troll", and no overall agreed upon exhaustive framework along the lines of what Sanovich et al. propose.

Prevalence of Phenomena

Another repeated concern across different reviews is the immediate jump in the scientific literature to measuring the effect of a phenomenon before having a good sense of its prevalence. To be clear, the academic incentives for scholarly publication—at least in the social sciences—lean toward establishing causal relationships, as opposed to counting exercises, so this development is understandable. That being said, smart public policy decisions depend on policymakers having a good understanding about the prevalence of activities in order to assess the costs and benefits of proposed policy changes. In particular, more information is needed concerning:

- The proportion of political conversations that occur online.
- The proportion of political conversations online that are cross-partisan
    - At the aggregate level (e.g., what is the average level of cross-partisan information to which individuals are exposed?).
    - At the individual level (e.g., what proportion of individuals find themselves in "echo chambers," and what are the characteristics distinguishing those who are exposed to cross-partisan information from those who are not?).
- The amount of actual online exposure to all of the different "disinformation" categories mentioned above, for a variety of different groups, including:
    - The modal social media user.
    - High-frequency social media users.
    - Politically interested citizens .
    - Liberals versus conservatives.
- The number of exclusively "fake news" websites producing political content and the size and composition of their audience.
- The quantity of political news stories produced by hyperpartisan versus partisan versus non-partisan news sources and the size and composition of their audience.
- The amount of disinformation shared by bots, cyborgs, and humans, and the size and composition of their audience. Regarding bots in particular, this includes
    - Human beings that bots are attempting to persuade/deceive.
    - Algorithms that bots are attempting to manipulate.
- The proportion of the top/trending stories on social media platforms that were originated/were amplified by bots.

Substantive Research Gaps:

1. The Effects of Exposure to Information and Disinformation Online

The overall consensus in empirical studies of information consumption on social media is that these platforms increase exposure to new information, either to ideologically diverse opinions (Bakshy et al. 2015) or misinformation (Fourney et al. 2017). What remains mostly unanswered, however, is how individuals react to this exposure process, and in particular the **causal mechanisms that may explain opinion change**. Three topics seem particularly important moving forward:

a) Offline Effects of Disinformation and Corrections: Updating versus Backlash

Bayesian theories of information processing would suggest that individuals update their political positions in response to new information, in a direction that is consistent with what they learned (Achen 1992; Bullock, 2009). However, many scholars have demonstrated the existence of **backlash or boomerang** effects that lead to individuals' reinforcing their previous positions (Lewandowsky et al. 2012; Nyhan & Reifler 2010), either due to motivated reasoning (Taber & Lodge 2006; Redlawsk 2002), varying interpretation of the same set of facts (Gaines et al. 2007), or other reasons.

Moreover, studies of the effectiveness of corrective information have found **widely varying results** (e.g., compare Nyhan & Reifler 2010, N.d., with Nyhan et al. N.d.).[45] Further research is needed to determine the conditions under which fact checking and other forms of corrective information are most effective, and can build on recent work about the importance of "unlikely" sources of corrections (Berinsky 2017) and the provision of alternative narratives (Nyhan & Reifler 2015). It is also necessary to consider the extent to which corrective information can generate lasting changes in belief, given the observed durability of misperceptions such as the birther myth.

Closely related, some studies find reduced belief polarization and, to a lesser extent, improvements in belief accuracy **in response to financial incentives** (Bullock et al. 2015, Prior et al. 2015). These studies suggest that survey measures of factual beliefs may include some measure of "partisan cheerleading." The extent to which these reveal insincere beliefs is unclear, however, given that strong accuracy incentives are not present in real-world politics (see Flynn et al. 2017 for an extended discussion of this point). In addition, a study of adherents to the Obama Muslim myth indicated that their beliefs appeared to be sincerely held (Berinsky 2018). Further research is needed to determine how to best dissuade partisan cheerleading, while maintaining realistic conditions, when measuring factual beliefs about controversial issues.

[45] For a nice recent summary, see the January 3, 2018, *Slate* cover story by Daniel Engber: https://slate.com/health-and-science/2018/01/weve-been-told-were-living-in-a-post-truth-age-dont-believe-it.html.

b) Online Effects of Exposure to Disinformation

There are notable scholarly disagreements regarding **the extent to which disinformation shared on social media has any effect on citizens' political beliefs** (Allcott & Gentzkow 2017; Guess et al. 2017) or the extent to which news consumption through this platform may be **exacerbating political polarization** (Barberá N.d.; Baskhy et al. 2015; Boxell et al. 2017; Flaxman et al. 2016; Peterson et al. N.d.). One potential explanation for this pattern of seemingly conflicting empirical evidence is that these studies rely on **different conceptualizations of "misinformation" and "polarization."** For example, sometimes the differences between rumors, false information, misleading information, and hyperpartisan information are blurry. Similarly, different characteristics of social media platforms may contribute to affective polarization but deactivate ideological polarization. While there are many studies defining each of these two terms (see e.g., Prior 2013 and Berinsky 2017), a clear gap in this growing literature on social media and politics is a **comprehensive meta-analysis of previous studies that takes into account the varying definitions of key terms**.

Moreover, there are a host of other important potential **effects of exposure to disinformation online,** beyond exacerbating political polarization and whether or not individuals believe in the veracity of the disinformation to which they have been exposed. Particularly in the aftermath of the 2016 U.S. election, research is needed on the effect of exposure to disinformation on **turnout** and **party/candidate choice***.* But we certainly would also like to know more about the possible effects of exposure to disinformation on **positions on issues,** as well as general **interest in politics** and **trust in political institutions and the media**.

Similar questions could also be asked about the effects of **exposure to uncivil conversations** online. In particular, do such interactions make people less likely to participate in political discussions generally (both offline and online), in cross-partisan conversations more specifically, or even change the online networks in which they are embedded (e.g., "defriending")?

c) Online Effects of Exposure to Bots and Trolls

Despite all the work that has been done in recent years in attempts to identify bots online and, more recently, to characterize their political activity, we have very little work to date on **the effects of exposure to online bots on human behavior**. For example, do humans update opinions and beliefs differently when (dis)information is provided by a bot, as opposed to by another human? Does this effect change if a bot is antagonistic versus friendly? Do bots manage to substantially increase the popularity metrics of disinforming and polarizing posts, and if so, under what conditions, and do users respond to these inflated metrics by becoming more likely to share the messages involved? And can humans even tell if they are interacting with bots as opposed to humans? Even less is known about trolls (the vast majority of research to date has focused on the actions and motivations of trolls, as opposed to the political impact on those being trolled).

To reiterate, we are still lacking in basic descriptive statistics in this area, such as how likely any given individual is to encounter a bot or troll in the course of their daily social media use or, put another way, the proportion of social media posts that average user encounters that are produced by bots or trolls.

***

Taken together, it seems clear that sorting out the relative impact of exposure to disinformation, online conversations, bots, and partisan echo chambers (as well as their relative prevalence online) ought to be a crucial prerequisite for anyone hoping to design policies to mitigate potential pernicious effects on politics from social media usage, as **different problems prompt different solutions**. For example, more ideological self-segregation online might reduce "uncivil interactions," which tend to occur among people who disagree with one another, but make it less likely that instances of disinformation are ever corrected. Conversely, enabling social media users to "fact check" their friends might reduce the amount of disinformation online in the short-term, but in the long-term lead to more ideologically segregated networks (if fact checking leads to "defriending" or if users mainly share fact checking information that is aligned with their political views [Shin & Thorson 2017]), thus making the spread of disinformation less likely to be impeded down the road. Alternatively, it may be the case that "fact checking" is only effective when it comes from "unlikely sources," and that cross-ideological questioning of the quality of information only increases belief in that information, which would suggest, perhaps paradoxically, that ideologically diverse communities are more likely to breed belief in the veracity of disinformation. Regardless of the specific forces that may be at work, the interrelatedness of these different factors (as illustrated in the introduction of this report in Figure 1) points to the importance of continued basic research as a way to insulate policy changes from unanticipated consequences.

2. Cross- and Multi-Platform Research

As should be evident from the preceding reviews, the vast majority of research on social media and politics to date has occurred using data from a single social media platform in a given research study. Yet there are, of course, numerous social media platforms, and many people have accounts on multiple platforms.[46] Moreover, the provision of political disinformation is clearly not limited to one or two particularly popular platforms; less popular platforms such as Reddit, 4chan, and 8chan may play outsized roles in this regard. Thus, research that explicitly **compares the prevalence of behavior and causal effects across different platforms** is especially needed. To give some examples:

- Are there **more civil (or uncivil) political conversations on some platforms than others**? If so, can we learn something about the design features of platforms (including news feed algorithms) that may or may not encourage civil political discussion?

[46] See in particular http://www.pewinternet.org/2016/11/11/social-media-update-2016/.

- Is there a **cross-platform pattern to the distribution of disinformation** online?
- Is there more **correction of disinformation on some platforms** than others? If so, why?
- Do **bots (or trolls) play different roles** on different platforms? And do they collaborate across platforms?
- What are the predictors of memes that emerge from the universe of memes out there to actually go viral?

a) The Facebook (and Google) Gap

As is now well known, when social media research involves data from only a single platform, more often than not that platform is Twitter. While there are very good reasons to justify using Twitter data to study politics, especially in the United States, it is, of course, not the most popular social media platform either in the United States or globally: that distinction belongs to Facebook. Simply put, if we want a better understanding of how social media usage is affecting U.S. politics along all the lines discussed in this report, **analysis of the effects of Facebook usage** needs to play a larger role in scientific studies.

It is worth noting, however, that to the extent that we want to better understand the effect of exposure to disinformation, we also would want to see more analysis of **the effects of exposure to information through Google searches**, which in turn would raise the importance of understanding of what exactly people see when using Google for search.

b) Links between Social Media and Traditional Media

Following a similar line of reasoning as in the previous section, we know that a non-trivial portion of the information shared on social media is content produced by traditional media sources; this is especially the case if we want to study the dissemination of information from partisan and hyperpartisan media sources. Further, we also know that traditional media sources often report on social media usage and include social media posts as part of news stories. Additional research on the relationship between **traditional media and social media** therefore appears important. Examples could include:

- The ways in which political rumors from social media migrate into traditional media stories.
- The effect on the life span of disinformation from being picked up by traditional media, or, conversely, the effect on the reach of disinformation produced by traditional media sources as a function of social media activity.
- The social media strategy of hyperpartisan media.
- The role of bots and trolls in manipulating newsfeed algorithms for political purposes.

3. Video

The vast majority of research surveyed in this report has focused on text as the source of disinformation. The future of disinformation, however, may be in images and, perhaps even more perplexingly, video.[47]

Systematic research centered on **audiovisuals,** rather than text, is therefore urgently needed to ascertain the effects of different types of visual and audiovisual messages on political polarization and disinformation. Various obstacles have hindered such research so far. First, the development of widely agreed upon concepts and measures of visual political content has been slower compared with political textual content (Griffin 2015). Second, storing and retrieving image and audiovisual content is more cumbersome than textual content. Third, analyzing audiovisual content is more complex because it conveys more information than text and is accordingly more difficult to code. Fourth, computational tools to automatically and reliably process and code images are still underdeveloped compared to those that treat textual content. Finally, not all the kinds of social media data that would be best suited to study these phenomena are publicly available to scholars.

4. Generalizability and Comparability of U.S. Findings

While there is a great deal in this report on the relationship between disinformation and political polarization and the quality of democracy in the United States, there are valuable scientific gains to be made from placing the findings from **U.S.-centered research in a more comparative context**.

For one, America's rigid two-party political system is fairly unique among advanced democracies. Most Western democracies, by contrast, do not have near-perfect two-party systems, have dealt with partisan media for decades, host relatively strong public service broadcasters, and feature institutional arrangements that do not require cross-party consensus for government to function. As a result, most non-American democracies have lived with polarization and disinformation for a long time, but are also experiencing disruptive social media influences similar to the U.S. political system. **Comparative research** would not only establish whether U.S. findings generalize to other countries, but **also to better understand what kinds of institutional and systemic conditions facilitate or hinder polarization and misinformation, thus yielding policy recommendations that can be relevant to the U.S. context**.

We also do not know nearly enough about how digital media can contribute to polarization and disinformation in more unstable, hybrid regimes and non-democratic regimes, where their disruptive role may arguably be larger than in established Western democracies because institutions may be weaker.

---

[47] See in particular https://www.nytimes.com/interactive/2018/01/02/technology/ai-generated-photos.html ; http://www.washington.edu/news/2017/07/11/lip-syncing-obama-new-tools-turn-audio-clips-into-realistic-video/

5. Different Effects on Different People: Ideological Asymmetries

While the previous discussion has focused on trying to ascertain the relationship between social media usage, political polarization, disinformation, and democratic quality, it is, of course, likely that **different individuals will react differently to the same stimuli;** in the language of social science research, this is known as "heterogeneous treatment effects."

There are limitless directions in which such research can be advanced, but for now we highlight one area that has come up in a few studies: **ideological asymmetry** (Adamic & Glance 2005; Barberá et al. 2015; Brady et al. 2017). There are two ways of thinking about this topic. The first is whether extremists tend to react differently than moderates; similarly, one could compare partisans to non-partisans. Alternatively, we can compare whether conservatives and liberals react differently. If either disinformation or political polarization affect conservatives and liberals differently, then this type of research would seem to be particularly important moving forward.

6. New Laws and Regulations

The governments of democratic countries such as Germany or Spain have taken steps to **regulate the content that can be shared on social media**, banning hate speech or misinformation and imposing fines on companies and users who post such content. The European Commission recently launched a Fake News initiative that is expected to recommend similar regulations at the E.U. level.[48] These decisions raise relevant normative and empirical questions regarding their desirability and effectiveness. When anyone can post anything under the protection of anonymity, **how do we strike a balance between freedom of speech and avoiding the free spread of misinformation**? And what are the long-term consequences of these measures regarding the pluralism of public debates? Are these measures thwarting exposure to diverse political views, which is generally considered a sign of democratic health? Further, what are the trade-offs associated with anonymity online, which may give people the opportunity to speak more freely about politics—especially in less open countries—but that also can provide an opportunity to engage in hateful and even threatening speech?

Finally, if states have the ability to regulate what content social media platforms may permit online, what does that imply about state control over the data provided by citizens to online platforms? Can states require the sharing of data by platforms with commercial competitors? For scholarly research? And what are the trade-offs in this regard with proprietary ownership of data, privacy concerns, and the "right to be forgotten"?

[48] https://ec.europa.eu/digital-single-market/en/fake-news

7. Bot Detection and Analysis

To the extent that bots play an important role in the spread of disinformation, then there is a great deal of work that remains to be done specifically in the field **of political bot detection and bot analysis**. As most extant work in this field has involved detecting and analyzing bots (1) over short periods of time, (2) in particular political contexts, and (3) on a single platform (Twitter), many important questions remain:

- What is the lifespan of an average bot or botnet?
- Can a bot detection model developed in one political context (country) be used to find political bots in another context?
- How long can a bot detection method developed in one country continue to accurately find bots even in that country? Put another way, what is the decay rate of bot detection methods?
- Can we build algorithms to detect bots on platforms beyond Twitter?
- Can we build algorithms to detect trolls?[49]
- Even if we can find bots and trolls, can we reliably code their political orientation once we find them?

8. Politicians and Disinformation

As is by now well known to anyone living in the United States, politicians have the ability to attract large numbers of followers on social media, and to utilize social media accounts to affect political discussion, and, accordingly, political polarization. Political elites also play a key role in linking political predispositions to factual beliefs and claims in controversial policy debates, but relatively little is known about how politicians and the media help disseminate myths or the process by which they become entrenched in partisan belief systems (Flynn et al. 2017). We also know relatively little about how to effectively change incentives for elites to dissuade them from promoting misinformation, though they may be more responsive to interventions than the public (see, e.g., Nyhan & Reifler 2015). Systematic research on the **role politicians play in spreading or debunking disinformation** would thus be extremely valuable moving forward.

[49] Although see King et al. 2017.

B. Key Data Needs

Executive Summary

This section presents a concise list of data needs for making progress on filling the research gaps outlined in the previous section. Data needs are divided into three categories: data that could be collected in the future by scholars with traditional funding, but that has not yet been collected; data that is prohibitively costly for individual scholars to collect, but that could be provided by a well-funded central research institute/data repository; and data that is not currently available for open scientific analysis due to the fact that it is the property of social media platforms and/or due to privacy concerns.

Philanthropic organizations are urged to consider the possibility of providing support for a managed data repository that would make social media data available for open scientific analysis, in conjunction with proper safeguards to protect individual privacy.

Introduction

We propose a threefold typology for categorizing data needs for advancing our understanding of social media, disinformation, and political polarization:

Type I Data**:** Data that could be collected by researchers with normal funding support, but that has not yet currently been collected.

Type II Data**:** Data that are prohibitively costly for most researchers to acquire, store, and access, but that could be maintained in a publicly accessible data repository (or is currently in a publicly accessible repository, but that is costly to access) and made available for open scientific analysis with proper safeguards. We propose that **establishing and funding such a repository** is an important role that could be played by philanthropic organizations, working in conjunction with social media platforms.

Type III Data: Data that are not currently accessible for open scientific research due to proprietary and/or privacy concerns. The most important Type III Data for answering the questions posed in this report is undoubtedly data from **Facebook**, due to its dominant role as the most popular social media platform among Americans, and its ownership of Instagram—the second most popular social network—and Facebook Messenger and WhatsApp, the most widely used mobile instant messaging platforms.

Type IIIa Data: Data that could be produced by researchers working in collaboration with social media platforms.

Type I Data Needs

Data that could be collected by researchers with normal funding support.

Validated measures of online information consumption: A key methodological limitation of past work in the study of misinformation on social media is the **lack of reliable and valid measures of online information consumption**. Scholars generally rely on self-reports (Allcott & Gentzkow 2017), indirect measures based on network structures (Bakshy 2015; Barberá et al. 2015), or web history and tracking logs (Gentzkow & Shapiro 2011; Guess 2014; Flaxman et al. 2016). However, even the best web tracking data cannot determine whether individuals actually consumed and understood the information to which they were exposed. While some of these methods are probably good enough to approximate news consumption, there is a clear need in the literature for a **systematic study that combines quantitative and qualitative methods to validate how individuals consume news on social media platforms**.

Survey data paired with social media data of survey respondents: This is particularly important for studying the question of *who* shares disinformation online. Self-reported measures on surveys of encountering disinformation are notoriously noisy, and social media data that can objectively record the sharing of disinformation often lack the necessary information to identify relevant demographic characteristics of those sharing the information. Pairing surveys with browser tracking or social media accounts of the user allows for rich demographic information on respondents, paired with actual objective measures of social media or news consumption data. One solution: Develop **replicable, transparent, and fair procedures that allow academics to match Facebook user-level data with their own survey or web tracking data**, while ensuring these users' privacy and right to decline providing consent.

Real-time, smartphone-based surveys of political conversations: The Facebook API is currently limited in the kinds of data it can provide to help with questions about political discussions. However, it might be feasible to take advantage of smartphone-based survey measurement techniques to collect **more immediate self-reported data on political conversations** that minimize error compared to current practices (Ohme et al. 2016).

Validated measures of "affective polarization": While most scholars agree that affective polarization is on the rise in most developed democracies, the evidence regarding how it varies across countries is not as clear, due in part to the complexity of developing comparable measures of affective polarization that capture the same concept in different contexts. Country-level characteristics, such as the structure of party competition or the varying relevance of social cleavages, can make it difficult to identify, for example, which are the relevant in-groups and out-groups. A key methodological and data access gap in this literature is a **survey instrument of affective polarization that is validated using behavioral measures, and provides a comparable metric across countries.**

"Ground Truth" examples of bots and trolls: One of the challenges of identifying bots and trolls is actually having "ground truth" of true positives needed for training machine learning models when the creators of bots or the actual trolls prefer to remain anonymous. (True negatives are much easier to find.) Such data sets have sometimes been created from leaked information, but that approach leaves researchers vulnerable to manipulation from deliberate leaks of false information. One potential avenue for moving forward would be for researchers developing algorithms for detecting bots and trolls to collaborate with ethnographic researchers to jointly try to cooperate with actual trolls or producers of bots. Such a project would, of course, require serious thought and ethical consideration.

Troll detection algorithms: Building on algorithmic approaches to bot detection and the existing impressive ethnographic research on trolls, future work could attempt to develop similar **troll-detection techniques,** and use these to examine troll behavior systematically over time.

Experimental work related to bots or tolls: This could include both lab experiments or "field" experiments in actual online environments and/or social media platforms.

Comprehensive data linking political elites with propagation of disinformation: This could include, for example, a data set of any category of disinformation that is present in the Twitter and Facebook accounts of political elites. Collecting this data would be a two-step process of scraping the relevant pages and then searching for known sources of disinformation.

Cross-national dataset of legal restrictions on posting disinformation: This type of data collection could vary both cross-sectionally (across countries) and over time. It might also be useful to contrast restrictions on speech online with those offline.

Type II Data Needs

Data that is potentially available, but prohibitive costly—in terms of funding, time, or start-up costs—for most researchers to collect on their own or even in small groups.

An archive of images and video related to disinformation, video propaganda, etc.: Collecting video and images is something most researchers can do. However, collecting, storing, and accessing large numbers of images and videos can be challenging. There are very large computing and start-up costs associated with analyzing images and especially video—this is not something most social scientists are trained to do. A publicly accessible archive **with preprocessed images and video,** including thorough metadata, would be extremely valuable. Such an archive, however, would raise a number of challenges. One would be to prevent it from becoming a repository of hateful and offensive images that could be used for malicious purposes. Another would be the need to avoid causing any harm to the individuals featured in the images and videos.

Real-time data on emerging disinformation: Studying the impact of misinformation on social media at scale requires data on its prevalence. One approach would be to try to use machine learning methods to automatically detect "fake news" stories. While these types of automatic classification tools are generally accurate at distinguishing hard-news versus soft-news stories, they may not perform as well when the classification task is identifying misinformation. It is even possible that these methods will never work, given that humans are often unable to make such distinctions. One alternative solution could be the **real-time development of a crowdsourced list of stories that may be considered as false or misleading**, along with a score based on multiple human annotators.

Comprehensive data on individual cross-platform news consumption: The emergence of social media platforms has contributed to a trend of growing media fragmentation (Prior 2007). To obtain a comprehensive view of citizens' news consumption, it is more important than ever to **measure such fragmented media diet by combining data from multiple sources,** including offline media use tracking.[50] In the case of Facebook data, where survey research shows a large share of news consumption takes place (Mitchell et al. 2017), addressing this data access need will necessarily require joint efforts between academics and industry partners.

Free access to the Twitter archive: Twitter currently has a full archive publicly available (GNIP), but the prices for accessing this archive can be prohibitively expensive for academic research, suggesting the primary intended users of the archive are commercial firms. Making the **GNIP Twitter Archive** freely available for open scientific research would play a major role in removing barriers to entry for a wide range of research projects. This would be especially valuable for addressing the "prevalence of phenomena" research gaps identified in the previous section. It would also be very useful for the study of networks and network effects, as collecting data from Twitter's Streaming API does not provide the account IDs of friends and followers, requiring many time-consuming calls to Twitter's

[50] https://www.nytimes.com/2017/12/28/business/media/alphonso-app-tracking.html

Resting API to build up networks. One idea would be for a **consortium of philanthropic organizations to jointly fund and establish free access to the Twitter archive for scientific research, perhaps in collaboration with the Library of Congress**.[51]

Archiving and pre-processing other social media platforms: There are other social media platforms that have been noted as avenues for the spread of disinformation (e.g., Reddit, 8chan) where posts can be made anonymously and are accessible by the public. Collecting and preprocessing these data present non-trivial challenges in terms of storage and technical skills. Thus **a dedicated repository that archived open access social media platforms in a searchable database format** would like also open up many research opportunities.

Bot repositories: As more research teams attempt to detect and monitor political bot activity in increasingly more contexts, and with research suggesting the possibility of "bot recycling," a **centralized and searchable archive of suspect bot accounts** could prove valuable, although certainly serious attention would need to be paid to security and access issues.

Archiving the production of hyperpartisan media outlets: Assuming the question of defining "hyperpartisan media" could be satisfactorily addressed, a searchable database of stories produced by hyperpartisan media sources would be a useful starting point for researchers looking at the effects of such media.[52] The same could be said for foreign media outlets identified as purveyors of disinformation. In both cases, the idea would be to decrease the barriers to entry in studying these topics.

Replication versus proprietary data: One final point worth noting: There is a growing movement across all the social sciences for research to be more open and transparent.[53] Part of this process involves making data used in studies accessible for replication studies, which is increasingly becoming a prerequisite for publication in top journals. Social media platforms, on the other hand, often have strong restrictions about the manner in which data can be shared. To be clear, there are serious privacy concerns at play here, in addition to any proprietary data ownership issues. This is therefore a case where two legitimate sets of concerns are likely to collide with one another, with the potential to cause serious impediments for researchers studying the topics under review in this report. One potential idea: a **password-protected data repository for replication studies created in partnership with the platforms** where scholars wishing to replicate previous studies could access the necessary data, but would agree to a number of privacy-protecting conditions in return for access to the data**.** This does not necessarily have to be created

[51] See: https://www.theverge.com/2017/12/26/16819748/library-of-congress-twitter-archive-project-stalled. The Library of Congress has said they will no longer archive every tweet because they do not have the resources to do so. Perhaps this is an opportunity for a philanthropic organization?

[52] It is possible that such a project could be incorporated as a part of MediaCloud (https://mediacloud.org/), or perhaps already exists within that project.

[53] In political science in particular, see https://www.dartstatement.org/.

from scratch, and possibly could be integrated into an existing archive such as the Harvard Dataverse Network.[54]

[54] https://dataverse.harvard.edu/

Type III Data Needs

Data not currently available for open scientific research.

Access to Facebook data: Facebook continues to be by far the most popular social media platform both in the United States and globally,[55] and is used by large numbers of people to consume news. Simply put, without access to Facebook data, understanding of the spread of disinformation through social media will be incomplete. To be clear, a great deal can be learned through analyzing publically available data on sites like Twitter and Reddit. Further, there are serious privacy concerns that are driving Facebook's current policies on sharing data with academic researchers. Nevertheless, a great deal more could be learned about many of the topics contained in this report if **a system for sharing Facebook data with scientific researchers** could be developed and implemented.

Access to Google data: Although receiving much less attention than Facebook—and not directly the subject of this report—it is beyond question that search algorithms on sites such as Google also play an important role in how Americans consume news. Thus **a system for sharing Google data with scientific researchers** would also likely help us to better understand the online distribution of disinformation.

Other social media data: It is worth noting that other social media platforms are providing an increasing role in how Americans consume news, as detailed in a recent report by the Pew Research Center on "News Use across Social Media Platforms".[56] These include open platforms such as Reddit and YouTube, but also the less publicly accessible **Snapchat, Instagram,** and **WhatsApp**.

[55] http://www.pewinternet.org/2016/11/11/social-media-update-2016/

[56] http://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/

Type IIIa Data Needs

Data that could be collected in collaboration with platforms.

Randomized online field experiments: An important explanation for why we still know little about how misinformation may affect political beliefs is the difficulty in generating **high-quality experimental evidence on the effects of exposure to fake news**. Studies that measure exposure to fake news on social media necessarily suffer from self-selection bias. Results from lab or survey experiments do not easily generalize because they are conducted in artificial environments. Even studies exploiting longitudinal natural experiments as a source of exogenous variation (Boxell et al. 2017) need to deal with issues such as composition bias and the fact that social media platforms are in a constant state of evolution, making it difficult to study their effect on long-term changes in polarization. While serious ethical and IRB considerations would need to be addressed in any research design to ensure informed consent among participants, running **experiments on social media platforms** offers one of the most promising avenues for addressing a host of topics contained in these reports. This could include:

- Effects of exposure to various forms of disinformation.
- Effect of receiving disinformation from different senders, including close friends, "weak ties," and non-human (bot) sources.
- Effects of attempts to correct disinformation within social media platforms.
- Effects of "validation" (i.e., likes by other people, retweets) on the effectiveness of corrections of disinformation.

Section IV: Works Cited

Abramowitz, Alan I., and Kyle L. Saunders. (2008). "Is polarization a myth?." *The Journal of Politics* 70(2): 542-555.

Abramowitz, Alan I. and Steven Webster. (2016). "The rise of negative partisanship and the nationalization of U.S. elections in the 21st century." *Electoral Studies* 41: 12-22.

Achen, Christopher H. (1992). "Social psychology, demographic variables, and linear regression: Breaking the iron triangle in voting research." *Political behavior* 14 (3): 195-211.

Adamic, Lada and Glance, Natalie. (2005). "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog." *LinkKDD '05 Proceedings of the 3rd International Workshop on Link Discovery*: 36-43.

Ahler, Douglas J. (2014). "Self-Fulfilling Misperceptions of Public Polarization." *Journal of Politics* 76(3): 607-620.

Ahler, Douglas J. and Gaurav Sood. (N.d.). "The Parties in Our Heads: Misperceptions About Party Composition and Their Consequences." Forthcoming, *Journal of Politics*.

Aldrich, John H. and David W. Rohde. (2000). "The consequences of party organization in the House: The role of the majority and minority parties in conditional party government." In *Polarized politics: Congress and the president in a partisan era*, Jon R. Bond and Richard Fleisher, eds. CQ Press.

Allcott, Hunt and Matthew Gentzkow. (2017). "Social media and fake news in the 2016 election." *Journal of Economic Perspectives* 31(2):1–28, 2017.

Althaus, Scott L. (1998). "Information Effects in Collective Preferences." *American Political Science Review* 92(3): 545-558.

American Political Science Association. (1950). "Toward a more responsible two-party system." A report of the Committee on Political Parties of the American Political Science Association. *American Political Science Review* 44(3).

Ananyev, Maxim, and Anton Sobolev. (2017). "Fantastic Beasts and Whether They Matter: Do Internet 'Trolls' Influence Political Conversations in Russia?". Paper presented at Midwest Political Science Association Annual Meeting, April 6–9, 2017. Chicago, IL.

Anderson, Ashley A. and Huntington, Heidi E. (2017). "Social Media, Science, and Attack Discourse: How Twitter Discussions of Climate Change Use Sarcasm and Incivility." *Science Communication*. 39(5): 598-620.

ANES Guide to Public Opinion and Electoral Behavior (2015). "Important Difference in What Democratic and Republican Parties Stand For 1952-2012." Downloaded March 16, 2018 from http://www.electionstudies.org/nesguide/toptable/tab2b_4.htm.

Applebaum, Anne, Peter Pomerantsev, Melanie Smith, and Chloe Colliver. (2017). "'Make Germany Great Again': Kremlin, Alt-Right and International Influences in the 2017 German Elections." Institute for Strategic Dialogue and the Arena Project at the LSE's Institute of Global Affairs. http://www.isdglobal.org/wp-content/uploads/2017/12/Make-Germany-Great-Again-ENG-081217.pdf.

Arceneaux, K., & Johnson, M. (2015). "More a Symptom than a Cause." In *American Gridlock: The Sources, Character, and Impact of Political Polarization*, edited by J. Thurber and A. Yoshinaka. Cambridge University Press, 309-36.

Bail, C. A. (2016). "Emotional Feedback and the Viral Spread of Social Media Messages about Autism Spectrum Disorders." *American journal of public health*, *106*(7), 1173-1180.

Bakir, Vian, and Andrew McStay. (2017). "Fake News and the Economy of Emotions: Problems, Causes, Solutions." *Digital Journalism*. 5(10):1-22.

Bakshy, Eytan, Dean Eckles, Rong Yan, and Itamar Rosenn. (2012) "Social Influence in Social Advertising: Evidence from Field Experiments." *Proceedings of the 13th ACM Conference on Electronic Commerce*, 146-161.

Bakshy, Eytan, Solomon Messing, and Lada A. Adamic. (2015). "Exposure to Ideologically Diverse News and Opinion on Facebook." *Science* 348 (6239): 1130-1132.

Barberá, Pablo (N.d.). "How Social Media Reduces Mass Political Polarization. Evidence from Germany, Spain, and the US." *In progress manuscript.*

Barberá, Pablo, and Gonzalo Rivero. (2015). "Understanding the Political Representativeness of Twitter Users." *Social Science Computer Review* 33 (6): 712-729.

Barberá, Pablo, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. (2015). "Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?" *Psychological Science*. 26(10): 1531-1542.

Barberá, Pablo, Yannis Theocharis, Zoltan Fazekas and Sebastian Popa (N.d.). "The Dynamics of Citizen Incivility Towards Politicians." *In progress manuscript.*

Barnidge, Matthew. (2017). "Exposure to Political Disagreement in Social Media Versus Face-to-Face and Anonymous Online Settings." *Political Communication* 34 (2): 302-321.

Bartels, Larry M. (1996). "Uninformed Votes: Information Effects in Presidential Elections." *American Journal of Political Science* 40(1): 194-230.

Bartels, Larry M. (2002). "Beyond the Running Tally: Partisan Bias in Political Perceptions." *Political Behavior* 24(2): 117-150.

Baum, M. A., & Groeling, T. (2009). "Shot by the Messenger: Partisan Cues and Public Opinion Regarding National Security and War." *Political Behavior*, *31*(2), 157-186.

Baum, Matthew A. and Tim J. Groeling (2009). *War stories: The causes and consequences of public views of war*. Princeton University Press

Baumgartner, Jody C., and Jonathan S. Morris. (2010). "MyFaceTube Politics: Social Networking Websites and Political Engagement of Young Adults." *Social Science Computer Review* 28 (1): 24-44.

Berger, J. (2011). "Arousal Increases Social Transmission of Information." *Psychological Science*, *22*(7), 891-893.

Berinsky, Adam J. (2017). "Rumors and Health Care Reform: Experiments in Political Misinformation." *British Journal of Political Science*. 47(2): 241-262.

Berinsky, Adam. (2012). "The Birthers are (Still) Back." YouGov.

Berinsky, Adam. (2018). "Telling the Truth about Believing the Lies? The Prevalence of Expressive Responding in Surveys." *Journal of Politics*. 80(1): 211-224.

Bessi, A., Zollo, F., Del Vicario, M., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). "Trend of Narratives in the Age of Misinformation." *PloS one*, 10(8), e0134641.

Bessi, Alessandro, and Emilio Ferrara. (2016). "Social Bots Distort the 2016 U.S. Presidential Election Online Discussion." *First Monday* 21 (11). https://doi.org/10.5210/fm.v21i11.7090.

Blair, Spencer, Jonathan A. Busam, Katherine Clayton, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, Brendan Nyhan, Morgan Sandhu, Rachel Sang, Rachel Scholz-Bright, Austin T. Welch, Andrew G. Wolff, and Amanda Zhou. (N.d.). "Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Banners in Reducing Belief in False Stories on Social Media." Manuscript in progress.

Bode, Leticia. (2016a). "Political News in the News Feed: Learning Politics from Social Media." *Mass Communication and Society* 19 (1): 24-48.

Bode, Leticia. (2016b) "Pruning the News Feed: Unfriending and Unfollowing Political Content on Social Media." *Research & Politics* 3 (3): 1-8.

Boididou, Christina, Stuart E. Middleton, Zhiwei Jin, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, and Yiannis Kompatsiaris. (2017). "Verifying Information with Multimedia Content on Twitter." *Multimedia Tools and Applications*, September, 1–27. https://doi.org/10.1007/s11042-017-5132-9.

Bolsen, T., & Druckman, J. N. (2015). Counteracting the politicization of science. *Journal of Communication*, *65*(5), 745-769.

Bond, Jon R., Richard Fleisher, and Jeffrey E. Cohen. (2015). "Presidential-Congressional Relations in an Era of Polarized Parties and a 60-Vote Senate." In *American Gridlock: The Sources, Character, and Impact of Political Polarization*, James A. Thurber and Antoine Yoshinaka, eds. Cambridge University Press.

Bonica, Adam, Nolan McCarty, Keith T. Poole, and Howard Rosenthal. (2013). "Why Hasn't Democracy Slowed Rising Inequality?" *Journal of Economic Perspectives* 27(3): 103–124.

Born, Kelly and Nell Edgington. 2017. "Analysis of philanthropic opportunities to mitigate the disinformation/propaganda problem", *Hewlett Foundation*, https://www.hewlett.org/wp-content/uploads/2017/11/Hewlett-Disinformation-Propaganda-Report.pdf

Boxell, Levi, Matthew Gentzkow, and Jesse M. Shapiro. (2017) "Greater Internet use Is Not Associated with Faster Growth in Political Polarization among US Demographic Groups." *Proceedings of the National Academy of Sciences* 114 (40): 10612–10617.

Brady, William, Julian Willis, John T. Jost, Joshua A. Tucker, and Jay Van Bavel. (2017). "Emotion Shapes Diffusion of Moral Content in Social Networks." *Proceedings of the National Academy of Sciences*. 114(28): 7313-7318. doi: 10.1073/pnas.1618923114

Brulle, R. J., Carmichael, J., & Jenkins, J. C. (2012). "Shifting Public Opinion on Climate Change: An Empirical Assessment of Factors Influencing Concern Over Climate Change in the US, 2002–2010." *Climatic Change*, *114*(2), 169-188.

Brundidge, Jennifer. (2010). "Encountering 'Difference' in the Contemporary Public Sphere: The Contribution of the Internet to the Heterogeneity of Political Discussion Networks." *Journal of Communication*. 60(4): 680-700.

Buckels, Erin, Paul D Trapnell, and Delroy L Paulhus. (2014). "Trolls Just Want to Have Fun." *Personality and Individual Differences*. 67: 97–102.

Bullock, John G. (2009). "Partisan Bias and the Bayesian Ideal in the Study of Public Opinion." *The Journal of Politics* 71 (3): 1109-1124.

Bullock, John G., Alan S. Gerber, Seth J. Hill, and Gregory A. Huber. (2015). "Partisan Bias in Factual Beliefs about Politics." *Quarterly Journal of Political Science* 10(4): 519-78.

Burkhardt, Joanna M. (2017). "History of Fake News." *Library Technology Reports*. 53(8): 5-2.

Byford, Jovan. (2011). Conspiracy theories: a critical introduction. Springer.

Cambria, Erik, Praphul Chandra, Avinash Sharma, and Amir Hussain. (2010). "Do not Feel the Trolls." *ISWC, Shanghai*.

Carmines, Edward G. and James A. Stimson. (1989). *Issue Evolution: Race and the Transformation of American Politics*. Princeton University Pres.

Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. (2011). "Information Credibility on Twitter." In Proceedings of the 20th International Conference on World Wide Web, 675–684. WWW '11. New York, NY, USA: ACM. https://doi.org/10.1145/1963405.1963500.

Chadwick, A., Vaccari, C. & O'Loughlin, B. (2017). "Do Tabloids Poison the Well of Social Media? Explaining Democratically-Dysfunctional News Sharing." Unpublished manuscript.

Chadwick, Andrew. (2010). "Britain's First Live Televised Party Leaders' Debate: From the News Cycle to the Political Information Cycle." *Parliamentary Affairs* 64(1): 24-44.

Chadwick, Andrew. (2011). "The Political Information Cycle in a Hybrid News System: The British Prime Minister and the 'Bullygate' Affair." *The International Journal of Press/Politics*. 16(1):3-29.

Chadwick, Andrew. (2013). *The Hybrid Media system: Politics and Power*. Oxford University Press.

Chavoshi Nikan, Hossein Hamooni, and Abdullah Mueen. (2016). "Identifying Correlated Bots in Twitter." Social Informatics. LNCS 10047, pp. 14–21.

Chen, Cheng, Kui Wu, Venkatesh Srinivasan, and Xudong Zhang. (2013). "Battling the Internet Water Army: Detection of Hidden Paid Posters." In 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), 116–20. https://doi.org/10.1145/2492517.2492637.

Chen, X., Sin, S. C. J., Theng, Y. L., & Lee, C. S. (2015, June). "Why do social media users share misinformation?" In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 111-114). ACM.

Chu, Zi, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. (2012). "Who is Tweeting on Twitter: Human, Bot, or Cyborg?" In IEEE Transaction on Dependable and Secure Computing 9(6): 811-824.

Clarke, Steve. (2007). "Conspiracy Theories and the Internet: Controlled Demolition and Arrested Development," *Episteme: A Journal of Social Epistemology*. 4(2):167-80.

Coleman, Gabriella. (2012). "Phreaks, Hackers, and Trolls: The Politics of Transgression and Spectacle." *The Social Media reader* 5: 99-119.

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). "Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data." *Journal of Communication 64*(2), 317-332.

Conover, Michael, Ratkiewicz, Jacob, Francisco, Matthew R., Gonçalves, Bruno, Menczer, Filippo, and Flammini, Alessandro. (2011). "Political Polarization on Twitter." *ICWSM* 133:89-96.

Cox, Gary W. and Mathew D. McCubbins. 2007. *Legislative Leviathan: Party Government in the House*. Cambridge University Press

Cresci, Stefano, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. (2016). "DNA-Inspired Online Behavioral Modeling and Its Application to Spambot Detection." *IEEE Intelligent Systems* 31 (5): 58–64.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). "The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race." *Proceedings of the 26th International Conference on World Wide Web Companion*: 963-972. Perth, Australia

Dalrymple, Kajsa E., and Dietram A. Scheufele. (2007). "Finally informing the electorate? How the Internet got people thinking about presidential politics in 2004." *International Journal of Press/Politics* 12 (3): 96-111.

Davis, Clayton, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. (2016). "BotOrNot: A System to Evaluate Social Bots." In Proceedings of the 25th International Conference on World Wide Web Companion, 273–74. https://doi.org/10.1145/2872518.2889302.

Davis, Richard. (2005). *Politics Online: Blogs, Chatrooms, and Discussion Groups in American Democracy*. Routledge.

del Vicario, Michela, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. (2016). "The Spreading of

Misinformation Online." *Proceedings of the National Academy of Sciences* 113 (3): 554-559.

Dewey, Caitlin. (2014). "This Is Not an Interview with Banksy." *Washington Post*. https://www.washingtonpost.com/news /the-intersect/wp/2014/10/21/this-is-not-an-interview- with-banksy/?tid=a inl&utm term=.8a9 5d83438e9

Dewey, Caitlin. (2016). "Facebook Fake-News Writer: I think Donald Trump is in the White House because of Me." *Washington Post*. https://www.washingtonpost.com/news/the-intersect/wp/2016/11/17/faceboo k-fake-news-writer-i-think-donald-trump-is-in-the-white-house-because-of-me/.

Diamond, Larry, Marc F. Plattner, and Christopher Walker. (2016). *Authoritarianism Goes Global: the Challenge to Democracy*. Johns Hopkins University Press.

DiFonzo, N., & Bordia, P. (2007). *Rumor Psychology: Social and Organizational Approaches*. American Psychological Association.

DiJulio, Bianca, Mira Norton, and Mollyann Brodie. (2016). "Americans' Views on the U.S. Role in Global Health." Kaiser Family Foundation.

Dimitrova, Daniela V., Adam Shehata, Jesper Strömbäck, and Lars W. Nord. (2014). "The Effects of Digital Media on Political Knowledge and Participation in Election Campaigns: Evidence from Panel Data." *Communication Research* 41 (1): 95-118.

Druckman, J. N., Peterson, E., & Slothuus, R. (2013). "How Elite Partisan Polarization Affects Public Opinion Formation." *American Political Science Review*, *107*(1), 57-79.

Duggan, Maeve, and Aaron Smith. (2016) "The Political Environment on Social Media." *Pew Research Center*.

Edwards, Chad, Autumn Edwards, Patric Spence, and Ashleigh Shelton. (2014). "Is That a Bot Running the Social Media Feed? Testing the Differences in Perceptions of Communication Quality for a Human Agent and a Bot Agent on Twitter." *Computers in Human Behavior* 33 (April): 372–76.

Eldridge II, Scott. (2017). *Online Journalism from the Periphery: Interloper Media and the Journalistic Field.* Routledge.

Eveland Jr., William P., Morey, Alyssa C., and Hutchens, Myiah J. (2011). "Beyond Deliberation: New Directions for the Study of Informal Political Conversation from a Communication Perspective." *Journal of Communication*. 61(6): 1082-1103.

Everett, Richard M., Jason R.C. Nurse, and Arnau Erola. (2016). "The Anatomy of Online Deception: What Makes Automated Text Convincing?" Proceedings of the 31st Annual ACM Symposium on Applied Computing, pp. 1115-1120.

Faris, Robert, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. (2017). "Partisanship, Propaganda, and Disinformation: Online Media and the 2016 US Presidential Election." Berkman Klein Center for Internet & Society Research Paper.

Fedor, Julie, and Rolf Fredheim. (2017). "'We Need More Clips About Putin, and Lots of Them:' Russia's State-Commissioned Online Visual Culture." *Nationalities Papers* 45 (2):161–81.

Ferrara, Emilio, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. (2016). "The Rise of Social Bots." *Communications of the ACM* 59 (7):96–104. https://doi.org/10.1145/2818717.

Ferrara, Emilio. (2017). "Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election." In progress manuscript.

Finn, Samantha, Panagiotis Takis Metaxas, and Eni Mustafaraj. (2014). "Investigating Rumor Propagation with TwitterTrails." arXiv:1411.3550. https://arxiv.org/pdf/1411.3550.pdf

Finnegan, William. (2016). "Donald Trump and the 'Amazing' Alex Jones," *The New Yorker*, http://www.newyorker.com/news/daily-comment/donald-trump-and-the-amazing-alex-jones.

Fiorina, Morris P. and Samuel J. Abrams. 2008. "Political Polarization in the American Public." *Annual Review of Political Science* 11: 563-588.

Flaxman, Seth, Sharad Goel, and Justin M. Rao. (2016). "Filter Bubbles, Echo Chambers, and Online News Consumption." *Public Opinion Quarterly* 80 (1): 298-320.

Fletcher, Richard, and Rasmus Kleis Nielsen. (2017). "Are People Incidentally Exposed to News on Social Media? A Comparative Analysis." *New Media & Society*, forthcoming.

Flynn, D.J., Brendan Nyhan, and Jason Reifler. (2017). "The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics." *Advances in Political Psychology* 38(S1): 127-150.

Foos, Florian, Lyubomir Kostadinov, Nikolay Marinov, and Frank Schimmelfennig.(N.d.). "Does Social Media Promote Civic Activism? A Field Experiment with a Civic Campaign." *In progress manuscript.*

Forelle, Michelle C., Philip N. Howard, Andrés Monroy-Hernández, and Saiph Savage. (2015). "Political Bots and the Manipulation of Public Opinion in Venezuela." In progress manuscript.

Fourney, Adam, Miklos Z. Racz, Gireeja Ranade, Markus Mobius, and Eric Horvitz. (2017). "Geographic and Temporal Trends in Fake News Consumption During the 2016 US Presidential Election." In *CIKM*, vol. 17, pp. 6-10. 2017.

Frankovic, Kathy. (2016). "Belief in Conspiracies Largely Depends on Political Identity." YouGov.

Frankovic, Kathy. (2017). "Republicans See Little Need for the Russia Investigation." YouGov.

Fredheim, Rolf. (2017). "Robotrolling." NATO Strategic Communications Centre of Excellence. https://www.stratcomcoe.org/robotrolling-20171.

Freeder, Seth. (N.d.). "Malice and Stupidity: Outgroup Motive Attribution and Affective Polarization." Manuscript in progress.

Friggeri, Adrien, Lada A. Adamic, Dean Eckles, and Justin Cheng. (2014). "Rumor Cascades." In *Proceedings of the International Conference on Weblogs and Social Media*.

Fritz, Ben, Bryan Keefer, and Brendan Nyhan. (2004). *All the President's Spin: George W. Bush, the Media, and the Truth*. Simon and Schuster.

Gaines, Brian J., James H. Kuklinski, Paul J. Quirk, Buddy Peyton, and Jay Verkuilen. (2007). "Same Facts, Different Interpretations: Partisan Motivation and Opinion on Iraq." *The Journal of Politics* 69 (4): 957-974.

Galinsky, A. D., & Moskowitz, G. B. (2000). "Perspective-Taking: Decreasing Stereotype Expression, Stereotype Accessibility, and In-Group Favoritism." *Journal of Personality and Social Psychology*, *78*(4), 708.

Garrett, R. K., Gvirsman, S. D., Johnson, B. K., Tsfati, Y., Neo, R., & Dal, A. (2014). "Implications of Pro-and Counterattitudinal Information Exposure for Affective Polarization." *Human Communication Research 40*(3), 309-332.

Garrett, R. K., Weeks, B. E., & Neo, R. L. (2016). "Driving a Wedge between Evidence and Beliefs: How Online Ideological News Exposure Promotes Political Misperceptions." *Journal of Computer-Mediated Communication 21*(5), 331-348.

Gentzkow, Matthew and Jesse M. Shapiro. (2011). "Ideological Segregation Online and Offline." *Quarterly Journal of Economics* 126(4): 1799-1839.

Gerber, Theodore P., and Jane Zavisca. (2016). "Does Russian propaganda work?." *The Washington Quarterly* 39(2): 79-98.

Giglietto, Fabio, Laura Iannelli, Luca Rossi, and Augusto Valeriani. (2016). "Fakes, News and the Election: A New Taxonomy for the Study of Misleading Information within the Hybrid Media System." In progress manuscript.

Gilens, Martin. (2001). "Political Ignorance and Collective Policy Preferences." *American Political Science Review* 95(2): 379-396.

Ginsberg, Benjamin and Martin Shefter. (1999). *Politics by Other Means: Politicians, Prosecutors and the Press from Watergate to Whitewater*. W.W. Norton.

Goel, S., Anderson, A., Hofman, J., & Watts, D. J. (2015). The structural virality of online diffusion. *Management Science*, *62*(1), 180-196.

González-Bailón, Sandra, Kaltenbrunner, Andreas, and Banchs, Rafael E. (2010). "The Structure of Political Discussion Networks: a Model for the Analysis of Online Deliberation." *Journal of Information Technology* 25(2): 230-243.

Gorodnichenko, Yuriy, Tho Pham, and Oleksandr Talavera. (2017). "Social Network, Sentiment and Political Outcomes: Evidence from #Brexit." https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=RESConf2017&paper_id=607.

Gottfried, Jeffrey, and Elisa Shearer. (2016). "News Use across Social Media Platforms 2016." *Pew Research Center*. http://www.journalism.org/2016 /05/26/news-use-across-social- media-platforms-2016.

Granovetter, Mark S. (1973). "The Strength of Weak Ties." *American journal of sociology* 78 (6): 1360-1380.

Griffin, M. (2015). "Visual Communication." In *The International Encyclopedia of Political Communication, e*dited by Gianpietro Mazzoleni. John Wiley & Sons.

Grimme, Christian, Mike Preuss, Lena Adam, and Heike Trautmann. (2017). "Social Bots: Human-Like by Means of Human Control?" *Big Data* 5 (4):279–93.

Groshek, Jacob, and Daniela Dimitrova. (2011). "A Cross-Section of Voter Learning, Campaign Interest and Intention to Vote in the 2008 American Election: Did Web 2.0 Matter." *Communication Studies Journal* 9 (1): 355-375.

Guess, Andrew M. (2014). "Measure for Measure: An Experimental Test of Online Political Media Exposure." *Political Analysis* 23 (1): 59-75.

Guess, Andrew M. (N.d.). "Media Choice and Moderation: Evidence from Online Tracking Data." Manuscript in progress.

Guess, Andrew, Benjamin Lyons, Brendan Nyhan, and Jason Reifler. (2017). "Avoiding the Echo Chamber about Echo Chambers: Why Selective Exposure to Congenial Political News is Less Prevalent than You Think." Knight Foundation report.

Guess, Andrew, Brendan Nyhan, and Jason Reifler. (N.d.a). "Selective Exposure to Misinformation: Evidence from the Consumption of Fake News During the 2016 U.S. Presidential Campaign." *In progress manuscript.*

Guess, Andrew, Brendan Nyhan, and Jason Reifler. (N.d.b). "'You're Fake News!' The 2017 Poynter Media Trust Survey." Poynter Institute.

Guess, Andrew, Briony Swire, Adam Berinsky, John Jost, and Joshua Tucker. (2017). "Rumors in Retweet: Social Media and the Spread of Political Misinformation." In progress manuscript.

Gupta, Aditi, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. (2013). "Faking Sandy: Characterizing and Identifying Fake Images on Twitter during Hurricane Sandy." *World Wide Web Conference WWW 2013 Companion*.

Hart, William, Dolores Albarracín, Alice H. Eagly, Inge Brechan, Matthew J. Lindberg, and Lisa Merrill. (2009). "Feeling Validated ersus Being Correct: a Meta-Analysis of Selective Exposure to Information." *Psychological Bulletin* 135 (4): 555.

Hasell, A., & Weeks, B. E. (2016). "Partisan Provocation: The Role of Partisan News Use and Emotional Responses in Political Information Sharing in Social Media." *Human Communication Research*, *42*(4), 641-661.

Hassan, Naeemul, Afroza Sultana, You Wu, Gensheng Zhang, Chengkai Li, Jun Yang, and Cong Yu. (2014). "Data in, Fact out: Automated Monitoring of Facts by FactWatcher." *Proc. VLDB Endow.* 7(13): 1557–1560.

Heath, C., Bell, C., & Sternberg, E. (2001). "Emotional Selection in Memes: the Case of Urban Legends." *Journal of Personality and Social Psychology 81*(6), 1028.

Herring, Susan, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. (2002). "Searching for Safety Online: Managing 'Trolling' in a Feminist Forum." *The Information Society* 18(5): 371-384.

Hetherington, Marc J. (2001). "Resurgent Mass Partisanship: The Role of Elite Polarization." *American Political Science Review* 95(3): 619-631.

Higgin, Tanar. (2013). "FCJ-159/b/lack up: What Trolls Can Teach Us About Race." *The Fibreculture Journal*. 22(2013).

Higgins, Andrew, Mike McIntire, and G. J. Dance. (2016). "Inside a Fake News Sausage Factory: 'This Is All About Income'." *The New York Times*, November 25, 2016.

Ho, Shirley S. and McLeod, Douglas M. (2008). "Social-Psychological Influences on Opinion Expression in Face-to-Face and Computer-Mediated Communication." *Communication Research* 35(2): 190-207.

Hopkins, Daniel J., John Sides, and Jack Citrin. (N.d.). "The Muted Consequences of Correct Information About Immigration." In progress manuscript.

Howard, Philip N. (2006). *New Media Campaigns and the Managed Citizen*. Cambridge University Press.

Howard, Philip N. and Bence Kollanyi. (2016). "Bots, #Strongerin, and #Brexit: Computational Propaganda During the UK-EU Referendum." In progress manuscript.

Howard, Philip N., Gillian Bolsover, Bence Kollanyi, Samantha Bradshaw, and Lisa-Maria Neudert. (2017). "Junk News and Bots during the US Election: What Were Michigan Voters Sharing Over Twitter?." *Project on Computational Propaganda.*

Howard, Philip, and Muzammil Hussain. (2013). *Democracy's Fourth Wave?: Digital Media and the Arab Spring*. Oxford University Press.

Huckfeldt, Robert and Sprague, John. (1995). *Citizens, Politics and Social Communication: Information and Influence in an Election Campaign.* Cambridge University Press.

Huckfeldt, Robert, Johnson, Paul E., and Sprague, John. (2004). *Political Disagreement: The Survival of Diverse Opinions Within Communication Networks.* Cambridge University Press.

Huckle, Steve, and Martin White. (2017). "Fake News: A Technological Approach to Proving the Origins of Content, Using Blockchains." *Big Data* 5 (4):356–71. https://doi.org/10.1089/big.2017.0071.

Huddy, Leonie, Lilliana Mason, and Lene Aaroe. (2015). "Expressive Partisanship: Campaign Involvement, Political Emotion, and Partisan Identity." *American Political Science Review* 109(1): 1-17.

Iyengar, Shanto and Sean J. Westwood. (2015). "Fear and Loathing across Party Lines: New Evidence on Group Polarization." *American Journal of Political Science* 59(3): 690-707.

Iyengar, Shanto, and Kyu S. Hahn. (2009). "Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use." *Journal of Communication* 59 (1): 19-39.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. (2012). "Affect, Not Ideology: A Social Identity Perspective on Polarization." *Public Opinion Quarterly* 76(3): 405-431.

Iyengar, Shanto, Kyu S. Hahn, Jon A. Krosnick, and John Walker. (2008). "Selective Exposure to Campaign Communication: The Role of Anticipated Agreement and Issue Public Membership." *Journal of Politics* 70 (01): 186-200.

Jacobs, Lawrence R., Cook, Fay Lomax, and Delli Carpini, Michael X. (2009). *Talking Together: Public Deliberation and Political Participation in America*. University of Chicago Press.

Jang, S. Mo, Lee, Hoon, and Park, Yong Jin. (2014). "The More Friends, the Less Political Talk? Predictors of Facebook Discussions Among College Students." *Cyberpsychology, Behavior, and Social Networking.* 17(5): 271-275.

Jay, Martin. (2010). *The Virtues of Mendacity: On Lying in Politics*. University of Virginia Press.

Jiang, L., Miao, Y., Yang, Y., Lan, Z., & Hauptmann, A. G. (2014). "Viral Video Style: A Closer Look at Viral Videos on YouTube." In *Proceedings of International Conference on Multimedia Retrieval*.

Jin, Zhiwei, Juan Cao, Yongdong Zhang, and Jiebo Luo. (2016). "News Verification by Exploiting Conflicting Social Viewpoints in Microblogs." *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, pp.2972-2978.

Jin, Zhiwei, Juan Cao, Yu-Gang Jiang, and Yongdong Zhang. (2014). "News Credibility Evaluation on Microblog with a Hierarchical Propagation Model*." IEEE International Conference on Data Mining*, pp.230-239.

Jones, John Ira. (2014). "Natural Born Shenanigans: How the Birther Movement Exacerbated Confusion Over the Constitution's Natural Born Citizen Requirement." *Regent University Law Review* 27(1): 1-23.

Kahan, Dan M., Ellen Peters, Erica Cantrell Dawson, and Paul Slovic. 2017. "Motivated Numeracy and Enlightened Self-Government." *Behavioural Public Policy* 1(1): 54-86

Karlsen, Geir. H. (2016). "Tools of Russian Influence: Information and Propaganda." In *Ukraine and Beyond* (pp. 181-208). Springer International Publishing.

Keane, J. (2013). *Democracy and Media Decadence.* Cambridge University Press.

Keller, Franziska, David Schoch, Sebastian Stier, and JungHwan Yang. (2017). "How to Manipulate Social Media: Analyzing Political Astroturfing Using Ground Truth Data from South Korea." ICWSM, 564–567.

Kenski, Kate, and Natalie Jomini Stroud. (2006). "Connections Between Internet Use and Political Efficacy, Knowledge, and Participation." *Journal of Broadcasting & Electronic Media* 50 (2): 173-192.

King, Gary, Jennifer Pan, and Margaret E. Roberts. (2017). "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument." *American Political Science Review* 111 (3):484–501. https://doi.org/10.1017/S0003055417000144.

King, Gary, Jennifer Pan, and Margaret Roberts. (2013). "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107 (02):326–43.

Klofstad, Casey A. (2009). "Civic Talk and Civic Participation: The Moderating Effect of Individual Predispositions." *American Politics Research* 37 (5): 856-878.

Klofstad, Casey A., Sokhey, Anand Edward, and McClurg, Scott D. (2013). "Disagreeing about Disagreement: How Conflict in Social Networks Affects Political Behavior." *American Journal of Political Science*. 57(1): 120-134.

Kollanyi, Bence, Philip N. Howard, and Samuel C. Woolley. (2016). "Bots and Automation over Twitter during the First US Presidential Debate." *Project on Computational Propaganda.*

Kosloff, Spee, Jeff Greenbern, Tom Schmader, Tom Dechesne, and David Weise. (2010). "Smearing the Opposition: Implicit and Explicit Stigmatization of the 2008 U.S. Presidential Candidates and the Current U.S. President." *Journal of Experimental Psychology: General* 139(3): 383-398.

Kuklinski, James H., Paul J. Quirk, Jennifer Jerit, David Schwieder, and Robert F. Rich. (2000). "Misinformation and the Currency of Democratic Citizenship." *Journal of Politics* 62(3): 790-816.

Kumar, Srijan , Francesca Spezzano, and VS Subrahmanian. (2014). "Accurately Detecting Trolls in Slashdot Zoo via Decluttering." In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*, ASONAM '14: 188–195, Beijing, China.

Kunda, Ziva. (1990). "The Case for Motivated Reasoning." *Psychological Bulletin* 108(3): 480-498.

Kushin, Matthew J. and Kitchener, Kelin. (2009). "Getting Political on Social Network Sites: Exploring Online Political Discourse on Facebook." *First Monday* 14(11-2).

Labzina, Elena. (2017). "Rewriting Knowledge: Russian Political Astroturfing as an Ideological Manifestation of the National Role Conceptions." Presented at American Political Science Association, August 31—September 3, 2017. San Francisco, CA.

Ladd, Jonathan M. (2011). *Why Americans Hate the Media and How It Matters.* Princeton University Press.

Lankina, Tomila, and Kohei Watanabe (2017). "'Russian Spring' or 'Spring Betrayal'? The Media as a Mirror of Putin's Evolving Strategy in Ukraine" *Europe-Asia Studies* 69(10): 1526-1556, DOI: 10.1080/09668136.2017.1397603

Layman, Geoffrey C., Thomas M. Carsey, and Juliana Menasce Horowitz. (2006). "Party Polarization in American Politics: Characteristics, Causes, and Consequences." *Annual Review of Political Science* 9: 83-110.

Lelkes, Yphtach. (2016). "Mass Polarization: Manifestations and Measurements." *Public Opinion Quarterly* 80 (1): 392-410.

Levendusky, M. S. (2013). "Why do Partisan Media Polarize Viewers?" *American Journal of Political Science 57*(3), 611-623.

Levendusky, Matthew S., James N. Druckman, and Audrey McLain. (2016). "How Group Discussions Create Strong Attitudes and Strong Partisans." *Research & Politics* 3 (2): 1-6.

Levitan, Lindsey, and Julie Wronski. (2014) "Social Context and Information Seeking: Examining the Effects of Network Attitudinal Composition on Engagement with Political Information." *Political Behavior* 36 (4): 793-816.

Lewandowsky, Stephan, Ullrich KH Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. (2012). "Misinformation and its Correction: Continued Influence and Successful Debiasing." *Psychological Science in the Public Interest* 13 (3): 106-131.

Li, H., & Sakamoto, Y. (2014). "Social Impacts in Social Media: An Examination of Perceived Truthfulness and Sharing of Information. *Computers in Human Behavior 41*, 278-287.

Lytvynenko, Jane and Craig Silverman. (2017). "The Money Is Rolling in for Liberal Hyperpartisan Sites and It's Tearing Some of Them Apart." *BuzzFeed News*, May 5, 2017.

Maréchal, Nathalie. (2017). "Networked Authoritarianism and the Geopolitics of Information: Understanding Russian Internet Policy." *Media and Communication.* 5(1):29-41.

Margolin, Drew B., Aniko Hannak, and Ingmar Weber. (2017) "Political Fact-Checking on Twitter: When Do Corrections Have an Effect?" *Political Communication* 1-24.

Marwick, Alice, and Rebecca Lewis. (2017)."Media Manipulation and Disinformation Online." *Data & Society Research Institute*.

Mason, Lilliana. (2015). "'I Disrespectfully Agree': The Differential Effects of Partisan Sorting on Social and Issue Polarization." *American Journal of Political Science* 59(1): 128-145.

McCright, Aaron M. and Riley E. Dunlap. 2011. "The Politicization of Climate Change and Polarization in the American Public's Views of Global Warming, 2001–2010." *Sociological Quarterly* 52(2): 155–194.

McGranahan, Carole. (2017). "An Anthropology of Lying: Trump and the Political Sociality of Moral Outrage." *American Ethnologist*. 44(2): 243-248.

Mejias, Ulises A., and Nikolai E. Vokuev. (2017). "Disinformation and the Media: The Case of Russia and Ukraine." *Media, Culture & Society* 37(9):1027-1042.

Mele, Nicco, David Lazer, Matthew Baum, Nir Grinberg, Lisa Friedland, Kenneth Joseph, Will Hobbs, and Carolina Mattsson. (2017). "Combating Fake News: An Agenda for Research and Action." In progress manuscript.

Messing, Solomon, and Sean J. Westwood. (2014) "Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online." *Communication Research* 41 (8): 1042-1063.

Metaxa-Kakavouli, Danaë, and Nicolás Torres-Echeverry. (2017). "Google's Role in Spreading Fake News and Misinformation." https://papers.ssrn.com/abstract=3062984.

Metaxas, Panagiotis and Mustafaraj, Eni. (2012). "Social Media and the Elections." *Science*. 338(6106): 472–473.

Metaxas, Panagiotis Takis. (2010). "Web Spam, Social Propaganda and the Evolution of Search Engine Rankings." *Web Information Systems and Technologies* 45: 170–182.

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). "Social and Heuristic Approaches to Credibility Evaluation Online. *Journal of Communication 60*(3), 413-439.

Mihaylov, Todor, Georgi Georgiev, and Preslav Nakov. (2015)."Finding Opinion Manipulation Trolls in News Community Forums." In progress manuscript.

Miller, Blake and Mary Gallagher. (2017). "The Progression of Repression: When Does Online Censorship Move toward Real World Repression?". http://www.blakeapm.com/research/repression.

Miller, Blake. (2017). "Automated Detection of Chinese Government Astroturfers Using Network and Social Metadata". http://www.blakeapm.com/research/astroturfing.

Mitchell, Amy, Jeffrey Gottfried, Elisa Shearer, and Kristine Lu. (2017) "How Americans Encounter, Recall and Act Upon Digital News." *Pew Research Center*.

Muddiman, Ashley and Stroud, Natalie Jomini. (2017). "News Values, Cognitive Biases, and Partisan Incivility in Comment Sections." *Journal of Communication.* 67(4): 586-609.

Munger, Kevin. (N.d.). "Experimentally Reducing Partisan Incivility on Twitter." In progress manuscript. URL: http://kmunger.github.io/pdfs/jmp.pdf

Munger, Kevin, Patrick Egan, Jonathan Nagler, Jonathan Ronen, and Joshua Tucker (N.d.) "Political Knowledge and Misinformation in the Era of Social Media: Evidence from the 2015 U.K. Election." In progress manuscript. URL: http://kmunger.github.io/pdfs/uk_paper_KM_final_JN_rv12.pdf

Mutz, D. C. (2007). Effects of "in-your-face" television discourse on perceptions of a legitimate opposition. *American Political Science Review*, *101*(4), 621-635.

Mutz, D. C., & Goldman, S. K. (2010). Mass media. In J. F. Dovidio, M. Hewstone, P. Glick & V. M. Esses (Eds.), *e Sage handbook of prejudice, stereotyping, and discrimination* (pp. 241–257). Thousand Oaks, CA: Sage.

Mutz, Diana C. (2002 "The consequences of cross-cutting networks for political participation." *American Journal of Political Science* 46 (4): 838-855.

Mutz, Diana C. (2006). *Hearing the Other Side: Deliberative Versus Participatory Democracy*. Cambridge University Press.

Mutz, Diana C., and Martin, Paul S. (2001). "Facilitating Communication across Lines of Political Difference: The Role of Mass Media." *American Political Science Review.* 95(1): 97-114.

Newman, Nic, Richard Fletcher, Antonis Kalogeropoulos, David AL Levy, and Rasmus Kleis Nielsen. (2017) "Reuters Institute Digital News Report 2017."

news on Facebook." https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outpe

Nied, Conrad, Leo Stewart, Emma Spiro, & Kate Starbird. (2017). "Alternative Narratives of Crisis Events: Communities and Social Botnets Engaged on Social Media." In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 263-266).

Nithyanand, Rishab, Schaffner, Brian, and Gill, Phillipa. (2017). "Online Political Discourse in the Trump Era." In progress manuscript. URL: https://arxiv.org/pdf/1711.05303.pdf

Nocetti, J. (2015). "Contest and conquest: Russia and global internet governance." *International Affairs*, 91(1):111–130.

Nyhan, Brendan and Jason Reifler. (2010). "When Corrections Fail: The Persistence of Political Misperceptions." 2010. *Political Behavior* 32(2): 303-330.

Nyhan, Brendan and Jason Reifler. (2015). "The Effect of Fact-checking on Elites: A Field Experiment on U.S. State Legislators." *American Journal of Political Science* 59(3): 628-640.

Nyhan, Brendan and Jason Reifler. N.d. "Do People Actually Learn From Fact-Checking? Evidence from a longitudinal study during the 2014 campaign." Manuscript in progress.

Nyhan, Brendan, and Jason Reifler. (2015). "Displacing Misinformation about Events: An Experimental Test of Causal Corrections." *Journal of Experimental Political Science* 2(1): 81–93.

Nyhan, Brendan, Ethan Porter, Jason Reifler, and Thomas J. Wood. N.d. "Taking Corrections Literally But Not Seriously? The Effects of Information on Factual Beliefs and Candidate Favorability." Manuscript in progress.

Nyhan, Brendan. (2010). "Why the 'Death Panel' Myth Wouldn't Die: Misinformation in the Health Care Reform Debate." *The Forum* 8(1).

Nyhan, Brendan. (2017). "Why the Fact-Checking at Facebook Needs to Be Checked." *New York Times*, October 23, 2017.

Nyhan, Brendan. (2017b). "Norms Matter." *Politico Magazine*, September/October 2017.

Oentaryo, Richard, Arinto Murdopo, Philips Prasetyo, and Ee-Peng Lim. (2016). "On Profiling Bots in Social Media." *Social Informatics*. LNCS 10046, pp. 92–109.

Office Of The Director Of National Intelligence. (2017). "Intelligence Community Assessment: Assessing Russian Activities And Intentions In Recent Us Elections."

Ohme, Jakob, Albaek, Erik, and de Vreese, Claes H. (2016). "Exposure Research Going Mobile: A Smartphone-Based Measurement of Media Exposure to Political Information in a Convergent Media Environment." *Communication Methods and Measures*. 10(2-3): 135-148.

Oliver, J. Eric, and Thomas J. Wood. (2014). "Conspiracy theories and the paranoid style (s) of mass opinion." *American Journal of Political Science.* 58(4): 952-966.

Papacharissi, Zizi. (2004). "Democracy online: Civility, politeness, and the democratic potential of online political discussion groups." *New Media & Society*. 6(2): 259-283.

Papanastasiou, Yiangos. 2017. "Fake News Propagation and Detection: A Sequential Model." https://papers.ssrn.com/abstract=3028354.

Paquet-Clouston, Masarah, Olivier Bilodeau, and David Décary-Hétu. (2017). "Can We Trust Social Media Data? Social Network Manipulation by an IoT Botnet." SMSociety'17. https://dl.acm.org/citation.cfm?id=3097301.

Pariser, Eli. (2011) *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Parker, David C.W. and Matthew Dull. 2009. "Divided We Quarrel: The Politics of Congressional Investigations, 1947–2004." *Legislative Studies Quarterly* 34(3): 319-345.

Parker, David C.W. and Matthew Dull. 2013. "Rooting Out Waste, Fraud, and Abuse

Pasek, Josh, Tobias H. Stark, Jon A. Krosnick, and Trevor Tompson. 2015. "What motivates a conspiracy theory? Birther beliefs, partisanship, liberal- conservative ideology, and anti-Black attitudes." *Electoral Studies* 40: 482–489.

Pennycook, G., Cannon, T. D., & Rand, D. G. (2017). Prior exposure increases perceived accuracy of fake news, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2958246

Pennycook, Gordon, and David Rand. (2017). "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Stories Increases Perceived Accuracy of Stories Without Warnings." SSRN Scholarly Paper. https://papers.ssrn.com/abstract=3035384 .

Pennycook, Gordon, Tyrone D. Cannon, and David G. Rand. N.d. "Implausibility and Illusory Truth: Prior Exposure Increases Perceived Accuracy of Fake News but Has No Effect on Entirely Implausible Statements." Manuscript in progress.

Peterson, Erik, Sharad Goel, and Shanto Iyengar (N.d.) "Echo Chambers and Partisan Polarization: Evidence from the 2016 Presidential Campaign." *In progress manuscript*

Phillips, Whitney. (2011)." LOLing at tragedy: Facebook trolls, memorial pages and resistance to grief online." *First Monday*, *16*(12).

Phillips, Whitney. (2015). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. MIT Press.

Pierskalla, Jan H., and Florian M. Hollenbach. (2013) "Technology and collective action: The effect of cell phone coverage on political violence in Africa." *American Political Science Review* 107 (2): 207-224.

Pomerantsev, Peter and Michael Weiss. (2014). "The Menace Of Unreality: How The Kremlin Weaponizes Information, Culture, And Money." A Special Report Presented By The Interpreter, A Project Of The Institute Of Modern Russia.

Porup, J.M. "How Mexican Twitter Bots Shut Down Dissent." (2015). *Motherboard*, August 24, 2015. https://motherboard.vice.com/en_us/article/how-mexican-twitter-bots-shut-down- dissent.

Preoţiuc-Pietro, Daniel, Ye Liu, Daniel Hopkins, and Lyle Ungar. (2017) "Beyond binary labels: political ideology prediction of twitter users." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics,* vol. 1, pp. 729-740.

Prior, M. (2013). Media and political polarization. *Annual Review of Political Science*, *16*, 101-127.

Prior, Markus, Gaurav Sood, and Kabir Khanna. 2015. "You Cannot be Serious: The Impact of Accuracy Incentives on Partisan Bias in Reports of Economic Perceptions." *Quarterly Journal of Political Science* 10(4): 489–518.

Prior, Markus. (2005) "News vs. entertainment: How increasing media choice widens gaps in political knowledge and turnout." *American Journal of Political Science* 49 (3): 577-592.

Prior, Markus. 2007. *Post-Broadcast Democracy: How Media Choice Increases Inequality in Political Involvement and Polarizes Elections*. Cambridge, UK: Cambridge University Press.

Prior, Markus. (2013) "Media and political polarization." *Annual Review of Political Science* 16: 101-127.

Rachel Sang, Rachel Scholz-Bright, Austin T. Welch, Andrew G. Wolff, and Amanda Zhou. N.d. "Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Banners in Reducing Belief in False Stories on Social Media." Manuscript in progress.

Radziwill, Nicole M., and Morgan C. Benton. (2016). "Bot or Not? Deciphering Time Maps for Tweet Interarrivals." ArXiv:1605.06555 [Cs], May. http://arxiv.org/abs/1605.06555.

Ramsay, Clay, Steven Kull, Evan Lewis, and Stefan Subias. 2010. "Misinformation and the 2010 Election: A Study of the US Electorate." WorldPublicOpinion.org.

Ratkiewicz, Jacob, Michael Conover, Mark Meiss, Bruno Goncalves, Alessandro Flammini, and Filippo Menczer. (2011a). "Detecting and tracking political abuse in social media." In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media:* 297–304.

Ratkiewicz, Jacob, Michael Conover, Mark Meiss, Bruno Goncalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. (2011b). "Truthy: Mapping the Spread of Astroturf in

Microblog Streams." In *Proceedings of the 20th International Conference on the World Wide Web*: 249–252.

Redlawsk, David P. (2002) "Hot cognition or cool consideration? Testing the effects of motivated reasoning on political decision making." *Journal of Politics* 64 (4): 1021-1044.

Resnick, Paul, Samuel Carton, Souneil Park, Yuncheng Shen, and Nicole Zeffer. (2014). "RumorLens: A system for analyzing the impact of rumors and corrections in social media." In *Proc. Computational Journalism Conference*.

Richey, Mason. (2017)."Contemporary Russian revisionism: understanding the Kremlin's hybrid warfare and the strategic and tactical deployment of disinformation." *Asia Europe Journal*. 15(54). 1-13.

Roberts, Margaret. 2018. *Censored: Distraction and Delay Within China's Great Firewall*. Princeton, NJ: Princeton University Press.

Rogers, Katie, and Jonah Engel Bromwich. (2016) "The Hoaxes, Fake News, and Misinformation We Saw on Election Day." *The New York Times,* November 8, 2016.

Rogowski, Jon C. and Joseph L. Sutherland. 2016. "How Ideology Fuels Affective Polarization." *Political Behavior* 38(2): 485–508.

Rojecki, Andrew and Sharon Meraz. 2016. Rumors and factitious informational blends: The role of the web in speculative politics. *New Media & Society* 18(1): 25-43.

Rubin, Victoria. (2017). "Deception Detection and Rumor Debunking for Social Media." In *The SAGE Handbook of Social Media Research Methods*, edited by Luke Sloan and Anabel Quan-Haase, 342–64.

Rudat, A., Buder, J., & Hesse, F. W. (2014). Audience design in Twitter: Retweeting behavior between informational value and followers' interests. *Computers in Human Behavior*, *35*, 132-139.

Sanovich Sergey. (2017). *Computational Propaganda in Russia: The Origins of Digital Misinformation*. Report for the Project on Computational Propaganda, Oxford Internet Institute. http://comprop.oii.ox.ac.uk/publishing/working-papers/computational-propaganda-in-russia-the-origins-of-digital-misinformation/.

Sanovich, Sergey, Denis Stukal, and Joshua A. Tucker. (2018). "Turning the Virtual Tables: Government Strategies for Addressing Online Opposition with an Application to Russia." Forthcoming in *Comparative Politics*.

Schäfer, Fabian, Stefan Evert, and Philipp Heinrich. (2017). "Japan's 2014 General Election: Political Bots, Right-Wing Internet Activism, and Prime Minister Shinzō Abe's Hidden Nationalist Agenda." *Big Data* 5 (4):294–309. https://doi.org/10.1089/big.2017.0049.

Schaffner, Brian F. and Cameron Roche. 2016. "Misinformation and motivated reasoning: Responses to economic news in a politicized environment." *Public Opinion Quarterly* 81(1): 86-110.

Schmehl, Karsten, and Jane Lytvynenko. (2017). "7 Out Of The 10 Most Viral Articles About Angela Merkel On Facebook Are False." *BuzzFeed*. July 27, 2017. https://www.buzzfeed.com/karstenschmehl/top-merkel-news.

Schudson, M. (1997). "Dynamics of Distortion in Collective Memory." In *Memory Distortion: How Minds, Brains, and Societies Reconstruct the Past* (pp. 346-364). Harvard University Press.

Shane, Scott. (2017) "The fake Americans Russia created to influence the election." *The New York Times,* September 7, 2017.

Shao, Chengcheng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. (2016). "Hoaxy: A Platform for Tracking Online Misinformation." World Wide Web Conference WWW'16 Companion, 745-750.

Shao, Chengcheng, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. (2017). "The Spread of Fake News by Social Bots." ArXiv:1707.07592 [Physics], July. http://arxiv.org/abs/1707.07592.

Shapiro, Jacob N., and Nils B. Weidmann. (2015) "Is the phone mightier than the sword? Cellphones and insurgent violence in Iraq." *International Organization* 69 (2): 247-274.

Shearer, E., and J. Gottfried. (2017) "News Use Across Social Media Platforms 2017." *Pew Research Center.*

Shin, Jieun, Lian Jian, Kevin Driscoll, and François Bar. (2017) "Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction." *New Media & Society* 19 (8): 1214-1235.

Shin, Jieun, and Kjerstin Thorson. (2017). "Partisan Selective Sharing: The Biased Diffusion of Fact-Checking Messages on Social Media". *Journal of Communication.* 67(2): 233–255

Shore, Jesse, Jiye Baek, and Chrysanthos Dellarocas. (N.d.) "Network structure and patterns of information diversity on Twitter." *In progress manuscript*.

Shorey, S. and Howard, P. N. (2016). Automation, Algorithms, and Politics| Automation, Big Data and Politics: A Research Review. *International Journal of Communication*, *10*, 24.

Sides, John. 2016. "Stories or science? Facts, frames, and policy attitudes." *American Politics Research* 44(3): 387-414.

Siegel, Alexandra, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. N.d. ""Trumping Hate on Twitter? Online Hate Speech and White Nationalist Rhetoric in the 2016 US Election Campaign and its Aftermath." *In progress manuscript.*

Silverman, C. (2015). *Lies, damn lies and viral content*. Tow Center for Digital Journalism White Papers, available at https://academiccommons.columbia.edu/catalog/ac:vdncjsxkvx

Silverman, Craig and Jeremy Singer-Vine. (2016). "Most Americans who see fake news believe it, new survey says." https://www.buzzfeed.com/craigsilverman/fake-news-survey.

Silverman, Craig. (2016). "This analysis shows how fake election news stories outperformed real

Silverman, Craig. (2016). "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook." *BuzzFeed*. November 16, 2016. https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook.

Silverman, Craig. (2016). "This analysis shows how fake election news stories outperformed real news on Facebook." *BuzzFeed*, November 16, 2016.

Skjeseth, Heidi Taksdal. (2017)."All the president's lies: Media coverage of lies in the US and France." Reuters Institute for the Study of Journalism, University of Oxford.

Starbird, Kate, Jim Maddock, Mania Orand, Peg Achterman, P., and Robert Mason. (2014). "Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 Boston marathon bombing." iConference 2014 Proceedings.

Stenberg, G. (2006). Conceptual and perceptual factors in the picture superiority effect. *European Journal of Cognitive Psychology*, *18*(6), 813-847.

Stewart, Leo, Amer Arif, and Kate Starbird. (2018). "Examining Trolls and Polarization with a Retweet Network." Unpublished manuscript.

Stromer-Galley, Jennifer. (2002). "New voices in the public sphere: A comparative analysis of interpersonal and online political talk." *Javnost/The Public*. 9: 23-42.

Stromer-Galley, Jennifer. (2003). "Diversity of Political Conversation on the Internet: Users' Perspectives." *Journal of Computer-Mediated Communication*. 8(3).

Stroud, N. J. (2011). *Niche news: The politics of news choice*. Oxford University Press on Demand.

Stroud, Natalie Jomini, Muddiman, Ashley, and Scacco, Joshua M. (2017). "Like, recommend, or respect? Altering political behavior in news comment sections." *New Media & Society.* 19(11): 1727-1743.

Stroud, Natalie Jomini. 2008. "Media use and political predispositions: Revisiting the concept of selective exposure." *Political Behavior* 30 (3): 341–366.

Stukal, Denis, Sergey Sanovich, Richard Bonneau, and Joshua A. Tucker. (2017). "Detecting Bots on Russian Political Twitter." *Big Data* 5 (4):310–24. https://doi.org/10.1089/big.2017.0038.

Suárez-Serrato, Pablo, Margaret Roberts, Clayton Davis, and Filippo Menczer. (2016). "On the influence of social bots in online protests." In *International Conference on Social Informatics* (pp. 269-278). Springer International Publishing.

Subramanian, Samanth. (2017). "Inside the Macedonian Fake-News Complex." *Wired*. https://www.wired.com/2017/02/veles-macedonia-fake-news/

Suhay, Elizabeth, Emily Bello-Pardo, and Brianna Maurer. (2018) "The Polarizing Effects of Online Partisan Criticism: Evidence from Two Experiments." *The International Journal of Press/Politics* 23 (1): 95-115.

Sundar, S. S. (2008). The MAIN Model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 72–100). Cambridge, MA: The MIT Press.

Sunstein, Cass and Adrian Vermeule. (2009). "Conspiracy Theories: Causes and Cures." *Journal of Political Philosophy.* 17(2): 202–227.

Sunstein, Cass R. (2017) *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

Sydell, Laura. (2016). "We Tracked Down a Fake-News Creator in the Suburbs. Here's What We Learned." *National Public Radio*. http://www.npr.org/sections/alltechconsidered/2016/11/23/503146770/npr-finds-the-head-of-a-covert-fake-news-operation-in-the-suburbs.

Taber, Charles S. and Milton Lodge. 2006. "Motivated skepticism in the evaluation of political beliefs." *American Journal of Political Science* 50(3): 755-769.

Taber, Charles S., and Milton Lodge. (2006) "Motivated skepticism in the evaluation of political beliefs." *American Journal of Political Science* 50 (3): 755-769.

The ANES Guide to Public Opinion and Electoral Behavior. 2012. "Important Difference in What Democratic and Republican Parties Stand For 1952-2012." Downloaded December 6, 2017 from http://www.electionstudies.org/nesguide/toptable/tab2b_4.htm.

The Politics of House Committee Investigations, 1947 to 2004." *Political Research Quarterly* 66(3): 630-644.

Theocharis, Yannis, and Will Lowe. (2016) "Does Facebook increase political participation? Evidence from a field experiment." *Information, Communication & Society* 19 (10): 1465-1486.

Theocharis, Yannis, Pablo Barberá, Zoltán Fazekas, Sebastian Adrian Popa, and Olivier Parnet. (2016) "A bad workman blames his tweets: the consequences of citizens' uncivil Twitter use when interacting with party candidates." *Journal of communication* 66 (6): 1007-1031.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2387-2395).

Thomas, Kurt, Damon McCoy, Chris Grier, Alek Kolcz, and Vern Paxson. 2013. "Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse." In Proceedings of the 22Nd USENIX Conference on Security, 195–210. SEC'13. http://dl.acm.org/citation.cfm?id=2534766.2534784.

Thompson, Derek. (2013) "Upworthy: I thought this website was crazy, but what happened next changed everything." *The Atlantic*, November 14, 2013.

Thorson, Emily. N.d. “Identifying and Correcting Policy Misperceptions.” Manuscript in progress.

Timberg, Craig. (2016) "Russian propaganda effort helped spread ‘fake news’ during election, experts say." *Washington Post*, November 24, 2016.

Townsend, Tess. (2016.) “Meet the Romanian Trump Fan behind a Major Fake News Site. Inc.” *INC*. http://www. inc.com/tess-townsend/ending-fed-trump-facebook.html.

Trabelsi, Amine, and Osmar R. Zaiane. (2014). “Mining Contentious Documents Using an Unsupervised Topic Model Based Approach.” 2014 IEEE International Conference on Data Mining, pp. 550--559. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7023372

Treré, Emiliano. (2016). "The dark side of digital politics: Understanding the algorithmic manufacturing of consent and the hindering of online dissidence." *IDS Bulletin* 47.1.

Tsay, Brian. (2017). “Disaggregating Public Security Propaganda on Chinese Social Media.” Presented at American Political Science Association, August 31—September 3, 2017. San Francisco, CA.

Tucker, Joshua A., Yannis Theocharis, Margaret E. Roberts, and Pablo Barberá. (2017) "From liberation to turmoil: social media and democracy." *Journal of democracy* 28 (4): 46-59.

Vaccari, Cristian, and Augusto Valeriani. (2015). "Follow the leader! Direct and indirect flows of political communication during the 2013 Italian general election campaign." *New Media & Society.* 17(7): 1025-1042.

Vanderhill, Rachel. (2013). *Promoting authoritarianism abroad*. Lynne Rienner Publishers.

Vargo, Chris, Lei Guo, and Michelle Amazeen. (2017). “The Agenda-Setting Power of Fake News: A Big Data Analysis of the Online Media Landscape from 2014 to 2016.” *New Media & Society*, https://doi.org/10.1177/1461444817712086.

Varol, Onur, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. (2017). “Online Human-Bot Interactions: Detection, Estimation, and Characterization.” ArXiv:1703.03107 [Cs], March. http://arxiv.org/abs/1703.03107.

Verba, Sidney, Burns, Nancy, and Schlozman, Kay Lehman. (1997). “Knowing and Caring about Politics: Gender and Political Engagement.” *Journal of Politics*. 59(4): 1051-1072.

Walsh, Katherine Cramer. (2004). *Talking about Politics: Informal Groups and Social Identity in American Life*. University of Chicago Press.

Walther, Joe. (2011). “Theories of Computer-Mediated Communication and Interpersonal Relations.” In *The Handbook of Interpersonal Communication* (ed. Knapp, Mark L. and Daly, John A.), 443-479. SAGE Publications.

Warren, T. Camber. (2015) "Explosive connections? Mass media, social media, and the geography of collective violence in African states." *Journal of Peace Research* 52 (3): 297-311.

Way, Lucan. (2015). “The limits of autocracy promotion: The case of Russia in the ‘near abroad’”. *European Journal of Political Research*, *54*(4), 691-706.

Weeks, B. E. (2015). Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation. *Journal of Communication*, *65*(4), 699-719.

Weeks, B. E., & Garrett, R. K. (2014). Electoral consequences of political rumors: Motivated reasoning, candidate rumors, and vote choice during the 2008 US presidential election. *International Journal of Public Opinion Research*, *26*(4), 401-422.

Weeks, Brian E. (2015) "Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation." *Journal of Communication* 65 (4): 699-719.

Wei, Wei, Kenneth Joseph, Huan Liu, and Kathleen M. Carley. (2015). “The Fragility of Twitter Social Networks Against Suspended Users.” 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 9-16.

Westerman, D., Spence, P. R., & Van Der Heide, B. (2014). Social media as information source: Recency of updates and credibility of information. *Journal of Computer-Mediated Communication*, *19*(2), 171-183.

Wiggins, Bradley E. (2017). "Navigating an Immersive Narratology: Factors to Explain the Reception of Fake News." *International Journal of E-Politics.* 8(3): 16-29.

Wood, Michael J., Karen M. Douglas, and Robbie M. Sutton. (2012). "Dead and alive: Beliefs in contradictory conspiracy theories." Social Psychological and Personality Science 3(6): 767-773.

Wojcieszak, M., & Azrout, R. (2016). I Saw You in the News: Mediated and Direct Intergroup Contact Improve Outgroup Attitudes. *Journal of Communication*, *66*(6), 1032-1060.

Wojcieszak, M., & Kim, N. (2016). How to improve attitudes toward disliked groups: The effects of narrative versus numerical evidence on political persuasion. *Communication Research*, *43*(6), 785-809.

Wojcieszak, M., Azrout, R., Boomgaarden, H., Alencar, A. P., & Sheets, P. (2017a). Integrating Muslim immigrant minorities: the effects of narrative and statistical messages. *Communication Research*, *44*(4), 582-607.

Wojcieszak, M., Azrout, R., De Vreese, C. (2017b). Waving The Red Cloth: Media Coverage Of A Contentious Issue Triggers Polarization, *Public Opinion Quarterly*, https://doi.org/10.1093/poq/nfx040

Wojcieszak, Magdalena E. and Mutz, Diana C. (2009). “Online Groups and Political Discourse: Do Online Discussion Spaces Facilitate Exposure to Political Disagreement?” *Journal of Communication*. 59(1): 40–56.

Wojcieszak, Magdalena. (2010) "‘Don’t talk to me’: effects of ideologically homogeneous online groups and politically dissimilar offline ties on extremism." *New Media & Society* 12 (4): 637-655.

Woolley, Samuel C., and Philip N. Howard. (2017). “Computational propaganda worldwide: Executive summary.” *Project on Computational Propaganda*.

Woolley, Samuel. (2016). “Automating Power: Social Bot Interference in Global Politics,” First Monday 21(4).

Wyatt, R. O., Katz, E., & Kim, J. (2000). “Bridging the Spheres: Political and Personal Conversation in Public and Private Spaces.” *Journal of Communication*. 50(1): 71-92.

Xu, Jun-Ming, Xiaojin Zhu, and Amy Bellmore. (2012). “Fast learning for sentiment analysis on bullying.” In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*.

Xu, Zhi and Sencun Zhu. (2010). “Filtering offensive language in online communities using grammatical relations.” In *Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*.

Zangerle, Eva, and Günther Specht. (2017). “"Sorry, I was hacked" A Classification of Compromised Twitter Accounts.” Proceedings of the 2014 ACM Symposium on Applied Computing, pp. 587-593.

Zhang, Huiling, Abdul Alim, Xiang Li, My T. Thai, and Hien T. Nguyen. (2016). “Misinformation in Online Social Networks: Detect Them All with a Limited Budget.” *ACM Transactions on Information Systems* 34(3): 18:2-18:24.

Zhou, Lina, and Dongsong Zhang. (2008). “Following Linguistic Footprints: Automatic Deception Detection in Online Communication.” *Communications of the ACM* 51(9): 119-122.

Ziegler, Charles E. (2017). "International dimensions of electoral processes: Russia, the USA, and the 2016 elections." *International Politics*: 1-18.

Zubiaga, A., & Ji, H. (2014). Tweet, but verify: epistemic study of information verification on twitter. *Social Network Analysis and Mining*, *4*(1), 163.